GRADUATE TEACHING ASSISTANTS' STATISTICAL KNOWLEDGE FOR

TEACHING

by

JENNIFER ANN NOLL

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY
in
MATHEMATICS EDUCATION

Portland State University
©2007

ACKNOWLEDGEMENTS

The completion of this dissertation would not have been possible without the support of a number of individuals. Foremost, I thank my research advisor, Karen Marrongelle, for her support, encouragement, and insightful feedback. I thank the other members of my dissertation committee for their support and feedback: Michael Shaughnessy, Luis Saldanha, Robert Fountain, and Swapna Mukhopadhyay. In particular, I thank Michael Shaughnessy. From our first time working together during my master's coursework in 1997, throughout our work on a National Science Foundation Grant, and during my dissertation study, Mike has always made himself available, provided insightful feedback, and been a source of support – I cannot thank Mike enough. Mike's ability to mentor graduate students, providing an intellectually demanding and supportive environment is an inspiration. I also wish to thank Luis Saldanha for providing insightful and substantive feedback on earlier versions of this document. I thank Bob Fountain, who is an amazing statistics teacher and someone who helped inspire my interest in statistics.

In addition, I thank Karen Noordhoff who provided support and feedback during the proposal stage of my dissertation work. Also, I thank John Caughman, Craig Swinyard, and Rachel Webb, who provided frequent assistance throughout my dissertation study. I cannot thank Craig enough for his professional and emotional support. I feel privileged to have Craig as a friend and colleague. Finally, I wish to thank my husband, Borg Norum, who is too wonderful to be real.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Increasingly, introductory college statistics is required for many majors and enrollment in introductory college statistics courses has been steadily increasing for the past decade (Luzter, Maxwell, & Rodi, 2000). Many college students will not have had any exposure to statistics in their K-12 curriculum because, although efforts have been underway to include statistics in the curriculum (National Council of Teachers of Mathematics [NCTM], 2000), only fairly recently have these efforts taken root. As more non-mathematics and non-statistics majors enroll in introductory statistics courses, teachers are faced with the challenge of teaching students that have increasingly diverse educational backgrounds. Students who enter introductory college statistics classes with an insufficient knowledge base are likely to experience difficulty comprehending the different statistical tests and procedures required in such a course. In addition, many undergraduate statistics courses are taught by mathematics or statistics graduate teaching assistants (TAs) (Luzter, Maxwell, & Rodi, 2000). While TAs teaching undergraduate statistics courses is not inherently problematic, it is not uncommon for TAs who majored in mathematics as undergraduates to enter graduate school having never taken a statistics course. Also, many TAs receive little preparation, orientation, or professional development before they begin their first teaching assignments (Belnap, 2005; Speer, 2001).

Undergraduate students with insufficient knowledge base and graduate TAs with insufficient background and experience represent two immediate challenges in the teaching of introductory statistics at colleges and universities. Although the focus of this research study is to consider in detail the second issue – that of TA background and experience, let me briefly address the first issue – that of student knowledge base. While the number of students taking introductory statistics is on the rise, the general population remains, for the most part, statistically illiterate[1] (Ben-Zvi & Garfield, 2004). The problem of statistical literacy has been addressed in recent years by NCTM (2000), which called for the inclusion of statistics in K-12 curriculum. The introduction of statistics into the K-12 mathematics curriculum should aid in the promotion of statistical literacy among the general population and help prepare students entering college for introductory statistics classes. However, it will take time before such curricular changes are implemented in all schools, and more time still for new generations of students to graduate having adequate statistics backgrounds.

At the college level, statisticians and statistics educators, such as Cobb and Moore (1997), and Cobb (1998), have called for reform in both the structure and content of introductory statistics courses. Cobb and Moore argue that statistics is fundamentally different than mathematics. Specifically, Cobb suggests that, "[s]tatistics is the science of data production and data analysis, and data analysis is an interpretive activity, albeit

---

[1] Statistical literacy is generally defined as the ability to organize statistical information, read information presented in tables and graphs, understand basic statistical concepts, and critically analyze statistical information typically found in newspapers and magazines. Thus, statistical illiteracy suggests the inability to process and understand basic statistical information. A more thorough explanation of statistical literacy will be provided in Chapter 2.

one that seeks to orient itself within a rigorous deductive framework" (Cobb, 1998, p. 3). Where mathematics can be abstracted to the point at which it is void of context, statistics cannot be stripped of its context without losing meaning. According to Cobb and Moore, this fundamental difference between statistics and mathematics has major implications for teaching statistics. In particular, introductory statistics should not be taught the way mathematics is often taught, devoid of context and focused on theorems, proofs, and procedures. Cobb and Moore argue that statistics must be taught with context in the forefront, and statistics courses should be focused on data production and analysis rather than abstracting theorems and statistical tests. But how are statistics courses being taught and by whom? This question points to the second fundamental challenge in teaching introductory college statistics courses – teacher knowledge and experience.

This study addresses the issue of teacher knowledge and experience. Specifically, the primary purpose of this study is to investigate TAs' statistical knowledge for teaching. In order to successfully implement reforms and teach introductory statistics courses in ways that serve the needs of the student body and the greater population, we need qualified teachers with an interest in stochastics[2] education and a sound understanding of the subject matter. Herein lies a very significant challenge for statistics educators for several reasons. First, many introductory statistics courses are taught by TAs with mathematics, rather than statistics backgrounds, who have little experience teaching, and more pressing priorities as graduate students. Second, there

---

[2] I use stochastics to refer to both probability and statistics, in the manner of Shaughnessy (1992).

tend to be few opportunities for TAs to participate in professional development for their teaching. Third, we know very little about beginning TAs' statistical knowledge for teaching. The most we can assume is that they are likely to teach in ways similar to how they were taught, and these methods may be at odds with current reform efforts.

1.1 Origin of Research Questions

There are many efforts underway to better understand student thinking, and to improve and reform the content of introductory college statistics courses (Chance, delMas, & Garfield, 2004; Cobb, 1993; Cobb & Moore, 1997; Tempelaar, 2002). For example, Chance, delMas, and Garfield (2004) investigated physical and computer simulations to provide students experience with authentic data production and analysis; Cobb (1993) reviewed a number of experimental statistics curricular projects. These research efforts point to ways that undergraduate statistics courses can be restructured in order to improve students' statistical thinking and development. The restructuring of undergraduate statistics courses requires new ways of teaching. Yet, little research has been done to address the preparation of teachers of college statistics. In order for statistics reform efforts to be successful, a greater understanding of teachers' conceptions is needed. Although this knowledge base is growing for K-12 teachers (Canada, 2004; Makar & Confrey 2004; Mickelson & Heaton, 2004; Watson, 2001; Watson & Moritz, 1997), it is lacking for teachers of undergraduate statistics. For instance, there are no studies about TAs' knowledge of statistics or how to teach statistics. Improvement and reform of undergraduate statistics courses can only evolve so far without a consideration of the role of TAs.

In thinking about ways to improve undergraduate statistics education, it is clear that the role of TAs is both a necessary topic to consider and a wide-open topic for research. Considering TAs' role in undergraduate statistics education, I wondered what statistical knowledge for teaching is necessary and sufficient for teaching in a manner that supports student learning and achievement. In order to narrow my field of study, I decided to concentrate on TAs' statistical knowledge for teaching sampling concepts[3]. My specific research questions are:

1. How do TAs understand sampling concepts? In particular,

    a. How do TAs conceptualize samples, the act of sampling and sampling distributions?

    b. How do TAs conceptualize the connections between sampling and statistical inference?

    c. How do TAs conceptualize the connections between probability and sampling concepts?

2. What knowledge base do TAs have of content and students? That is, what knowledge do TAs have of common student solution strategies or difficulties?

To be clear, all of these research questions address TAs' statistical knowledge for teaching with respect to the concepts of sampling. The goal of this research is to characterize TAs' statistical knowledge for teaching sampling processes by

_____

[3] The reason for choosing sampling concepts is discussed shortly.

constructing an explanation for how TAs reason about sampling, connections they draw between sampling and other areas of the statistics curriculum, and how they understand student thinking in this area. My focus on sampling concepts is intentional and significant. Sampling concepts are foundational and motivate statistical inference processes. Further, sampling concepts are abstract and difficult to mentally unpack, making them fertile ground for investigating TA reasoning in this domain. I elaborate further on the importance of studying sampling in the next section.

The research questions presented here aim to unite research on teacher knowledge, research on graduate teaching assistants, and research on probability and statistics education in the domain of sampling. It is reasonable to wonder if it makes sense to unite these seemingly distinct research areas and if this union has any importance or relevance to mathematics education. In the next section, I address the rationale for tying together research on undergraduate statistics education, in the domain of sampling with research on TAs' knowledge.

1.2 Rationale

1.2.1 Why Study Undergraduate Statistics Education?

Unfortunately, much of the research literature indicates that many adults are statistically illiterate and experience difficulty making informed decisions when confronted with quantitative information (Ben-Zvi & Garfield, 2004; The National Council on Education and the Disciplines [NCED], 2001). Issues of equity are intimately tied to statistical literacy, and statistical literacy is a necessary component for equal participation in a democratic society (NCED, 2001). Understanding public

issues, managing personal finances, and making personal health care decisions all require some level of statistical competence. In addition, many professions and higher paying jobs require statistical knowledge. At the post-secondary level more and more students are required to take introductory statistics in their degree programs. Enrollment in elementary statistics courses (non-calculus based) at four-year colleges and universities in the United States rose 18% from Fall of 1995 to Fall of 2000, and by 45% from 1990 levels (Luzter, Maxwell, & Rodi, 2000).

Undoubtedly the increased use of statistics in different careers, the increased enrollment in introductory statistics at the collegiate level, the increased use of statistics in today's media, and issues of equity constitute four major driving forces behind the mathematics education community's recent concern with the teaching and learning of probability and statistics. Ben-Zvi & Garfield (2004) suggest that, "[b]eing able to properly evaluate evidence (data) and claims based on data is an important skill that all students should learn as part of their educational programs" (p. 8). Such concerns over statistical literacy have prompted the NCTM (2000) to place increased attention on the promotion of statistical literacy and stochastics education in *The Principles and Standards for School Mathematics*. In fact, NCTM echoes Ben-Zvi and Garfield's argument:

> Statistics are often misused to sway public opinion on issues or to misrepresent the quality and effectiveness of commercial products. Students need to know about data analysis and related aspects of probability in order to reason statistically—skills necessary to becoming informed citizens and intelligent consumers (p. 47).

Yet, statistical ideas can be counterintuitive, students may have difficulty with the underlying mathematics, and students may rely on faulty intuitions when solving problems (Ben-Zvi & Garfield, 2004). Misconceptions in reasoning about probability and statistics are quite common even among those with considerable statistical training (Kahneman & Tversky, 1972; Konold, Pollatsek, Well, Lohmeier, & Lipson, 1993; Tversky & Kahneman, 1971). When one considers the increased use of statistics and data in today's society, the rise in statistics enrollment at colleges and universities, the importance of statistical literacy for equal participation in a democratic society, and the challenges of teaching statistics, it is evident that research efforts are critically needed in the area of stochastics education.

1.2.2 Why Focus on Sampling Concepts?

Statistical inference is the central focus for college-level introductory probability and statistics courses. Statistical inference is the process by which conclusions about a particular population are drawn from evidence provided by a *sample* of the population (Pfannkuch, 2005). Using statistical inference to make conclusions about different populations is commonplace in all fields today. For instance, statistical inference is used in medical science to make predictions about disease or surgeries, voter polls to make predictions about a candidate's success, and insurance to make predictions about accident rates.

Clearly statistical inference is an important skill for data-driven societies, and therefore is a key topic in introductory statistics courses. Yet, research suggests that

there are substantial gaps in students' understanding of statistical inference (Chance, delMas, & Garfield, 2004; Pfannkuch, 2005). Part of the difficulty lies in students' ability to make connections between probability models and statistical inference. In order for students to make salient connections between probability and statistical inference, they need a strong foundational understanding of distribution, variability, samples, and sampling distributions (Chance, delMas, & Garfield, 2004; Pfannkuch, 2005).

The discussion above suggests that an understanding of sampling concepts and processes is necessary for developing a robust understanding of statistical inference. Yet, sampling concepts can be quite difficult because they contain several layers of abstraction that must be understood before connections can be made between sampling and statistical inference techniques. Before discussing some of the difficulties in understanding sampling concepts, I briefly elaborate on what I mean by the terminology *sampling concepts and processes*. When I refer to sampling concepts and processes I am referring to samples, the act of sampling, sampling distributions, and/or measures of center (such as mean, median, and mode) and spread (such as range, variance, and standard deviation) associated with particular samples or sampling distributions. A *sample* is defined as a subset of a population. Statistical inferences made about a population are only reliable if the information obtained in the sample is representative of the population from which it was drawn. A typical way to collect representative samples is through the process of random sampling, meaning each element of the population has an equally likely chance of being included in the

sample. Also, smaller samples tend to be more variable (in the sense of spread around the mean) and thus have the potential to be less representative of the population. Thus, unpacking the concept of sample in relation to statistical inference entails knowledge of the definition, as well as understanding how to find representative samples and the consequences of obtaining biased samples.

The *distribution of a sample* represents the outcomes for a particular sample drawn from a population. For example, suppose a population consists of 200 introductory college statistics students from a large lecture hall course. Pick a random sample of 15 students from this class and note their scores on the midterm exam. The sample consists of 15 students' test scores. The graph shown in Figure 1 is a representation of the distribution of those 15 students' test scores.

Figure 1: Distribution of a Sample

Sample_of_15_Introductory_Statistics_Test_Scores
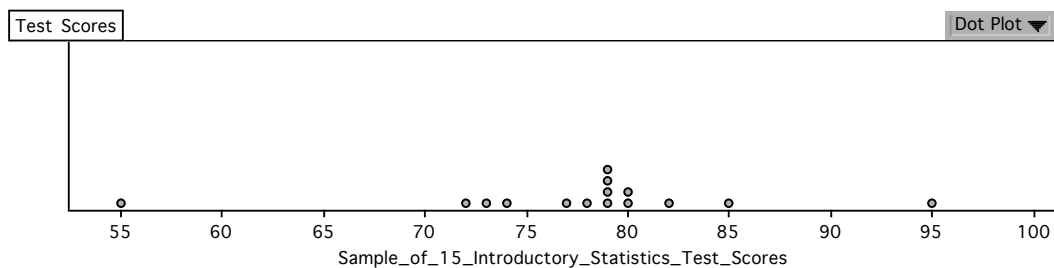
This information could be useful for finding the mean test score for the 15 students and/or a measure for how much each student deviated from the mean score. Thus, an understanding of, and an ability to compute, measures of center, shape, and spread is necessary in order to reason and draw inferences about the distribution of a sample.

Although statistical inferences are often based on information obtained from a single sample, the theory behind this process is based on the idea of repeated sampling

and the image of a sample as a particular case in a group of similar cases (Saldanha, & Thompson, 2003). Repeated sampling is necessary for the generation of a *sampling distribution*. A sampling distribution is the distribution of sample statistics (such as means, proportions, counts, etcetera) from all possible samples of size *n* drawn from the population. For instance, consider the example used in the preceding paragraph of the 200 introductory college statistics students. Suppose the sampling process is repeated over and over again – that is, random samples of size 15 are taken from this population. The mean test score for each sample is computed and the sample is replaced back into the population before sampling again. Once all possible combinations of 15 students are collected and their mean scores are marked, the distribution of those means represents the sampling distribution for this population of introductory college students' mean test scores.

Sampling distributions have two important properties: the mean of the sampling distribution is equal to the mean of the population, and the standard deviation of the sampling distribution is equal to the standard deviation of the population divided by the square root of the sample size. For the introductory statistics class example, one would not bother to find all possible sample test scores of size 15 from the introductory statistics class because the population mean and standard deviation (i.e., the mean test score and standard deviation for the class of 200 statistics students) can actually be computed. However, in most situations it would be impossible to compute the true population parameters. Instead, population parameters must be estimated by a single sample or several samples, so properties of sampling distributions are important

for understanding how those estimates are derived. Sampling distributions are inherently challenging because they require the act of repeated sampling and creating a distribution of sample statistics, such as the mean. Further, examining measures of center for a sampling distribution involves the added abstraction of computing the mean for a distribution of means, and examining standard deviation requires an understanding of how to compare standard deviations between samples and a recognition that the standard deviation of the sampling distribution is less than the standard deviation of the population.

Knowledge of repeated sampling, sampling distributions, and their properties sets the stage for understanding formal statistical inference procedures, such as computing confidence intervals. Confidence intervals provide an interval estimate for a population parameter, have a measure of reliability associated with their estimate, and are computed on the information provided by a single sample. The reliability measure (confidence level) for the confidence interval is based on the act of repeated sampling and is defined as "the relative frequency with which the interval estimate encloses the population parameter when the estimator is used repeatedly a very large number of times" (McClave & Sincich, 2000, p. 282).

The preceding paragraphs provide an outline of: (a) how I use sampling terminology; (b) what distinctions exist between relevant terms; (c) the conceptual complexity of sampling; and, (d) the importance of sampling concepts for building the foundations of statistical inference. Without an understanding of the sampling process, the distinction between the distribution of a sample and a sampling distribution, and

variability within and between samples, it is unlikely that students will develop informal or formal understandings of statistical inference.

Given that an understanding of the concepts of sampling are foundational to understanding statistical inference and that these concepts are rather complex, it is important to investigate TAs' reasoning and understand how they unpack these concepts. Specifically, how do TAs specify the important aspects of sampling and in what ways do they connect sampling concepts to other aspects of the statistics curriculum? How do TAs articulate ideas of sampling and what do they know about student thinking in this area?

1.2.3 Why Study Mathematics Graduate Teaching Assistants?

Introductory college statistics is likely to be the first exposure many students have to statistics. Students will form their attitudes and beliefs about the use of statistics from these beginning courses. In addition, introductory statistics courses are a recruiting ground for future statisticians (Moore, 2005). Thus, these introductory courses serve a critical function. At many universities, TAs teach the bulk of the introductory statistics courses or teach recitation sections for large lecture classes. In fact, the 2000 CBMS survey revealed that in the United States, TAs taught 21% and 24% of elementary statistics students in Ph.D. granting statistics departments and mathematics departments, respectively. TAs taught 43% of introductory statistics sections consisting of less than 36 students (Lutzer, Maxwell, & Rodi, 2000). Further, many TAs will be future professors teaching the next generation of mathematicians,

statisticians, and mathematics teachers. Thus, TAs have the potential to play a critical role in undergraduate statistics education and the promotion of statistical literacy among college students.

As TAs serve a critical role in undergraduate statistics education, successful efforts toward reforming undergraduate statistics and mathematics courses cannot be achieved without the inclusion of TAs. Yet, research on TAs' knowledge of the teaching and learning of mathematics is sparse at best. Speer, Gutmann, & Murphy (2005) argue that teachers' teaching practices form early in their careers, and so there exists a need for research on TAs' knowledge, as well as the implementation of professional development programs for fostering good teaching practices among TAs. The importance of the introductory statistics course in promoting statistical literacy, the complex nature of sampling concepts, and the central role of TAs in the teaching of introductory statistics provides a strong case that research addressing TAs' statistical knowledge for teaching sampling concepts is an important topic for the mathematics education community to consider.

In summary, the elevated role that statistics is given in our increasingly data driven society makes it an important educational topic of research. Given that the sampling process is fundamental to basic statistical reasoning, it is important that this topic receive attention in the research literature. Furthermore, given that universities and colleges have an important role to play in educating students and preparing them for a variety of careers, it is natural to look toward improving college teaching. Studying

TAs' statistical knowledge for teaching of sampling processes is an important first step in making improvements in the teaching of undergraduate statistics.

1.3 Chapter Abstracts

In Chapter 2, I provide a review of the background theories and literature relevant to my research project. Chapter 2 is presented in three sections. In Section 2.1, I discuss my guiding philosophy of learning – radical constructivism, and I discuss the impact of this philosophy on my research. In Section 2.2, I review the research literature on stochastics education. Specifically, I discuss the constructs of *statistical thinking*, *statistical literacy*, and *statistical reasoning*. These constructs are essential features that serve to frame my discussion of statistical knowledge for teaching. Statistical literacy provides a picture of the statistical knowledge necessary for participation in a democratic society. Statistical thinking illuminates "normative" modes of statistical inquiry; that is, the type of thinking employed by statisticians in their work. Statistical reasoning provides insight into how individuals reason about statistics, and suggests models of developmental stages. In Section 2.3, I address research on teacher knowledge. There is little research on TAs, and no prior research on statistics TAs. However, I examine research on teacher knowledge, and frameworks for successful teaching and professional development. Research on teacher knowledge serves as a useful basis for investigating TAs' knowledge. Finally, I meld research on statistics education with research on teacher knowledge as a means for framing this research on statistical knowledge for teaching.

In Chapter 3, I provide a discussion of my research methodology, study design, and analysis procedures. This chapter is presented in five sections. In Section 3.1, I provide a general overview of my data collection methods and rationale for those methods. Specifically, I address how the different elements of Chapter 2 work together to frame my study. In Section 3.2, I discuss the design of the research instruments used in this study. In Section 3.3, I discuss participant selection methods. In Section 3.4, I discuss my data analysis methods. In Section 3.5, I address issues of validity in this study.

In Chapter 4, I present the first of three themes that emerged from this research. A major finding of this study is that TAs experienced difficulties reasoning about experimental data. Specifically, TAs appeared to experience tension between their knowledge of theoretical models and their expectations for experimental data. In this Chapter I discuss the nature and source of this tension and the ways in which TAs tried to resolve their tensions. The Chapter is presented in two sections. In Section 4.1, I provide a conceptual analysis and framework for TAs reasoning with respect to two related tasks – The Prediction and Real/Fake Tasks[4], which entail sampling from a known population, and examining the likelihood of four experimental sampling distributions. Also, I use this framework to discuss how the TAs in this study reasoned about these tasks. In Section 4.2, I provide a conceptual analysis and framework for TAs reasoning with respect to the Mystery Mixture Task[5], which entails making

---

[4] The Prediction and Real/Fake Tasks can be found in the appendix.
[5] The Mystery Mixture Task can be found in the appendix.

decisions from four empirically collected sampling distributions. I use this framework to discuss how the TAs in this study reasoned about that task. The conceptual frameworks developed in this chapter are end products to this study and suggest a model of how TAs may reason and think about experimental data in a sampling situation.

In Chapter 5, I present the second of three themes that emerged from this research. A second major finding of this study is that the TAs in this study appeared to reason about certain sampling and statistical inference tasks along a spectrum, ranging from no connection to stronger connections of a long-term relative frequency interpretation of probability. Also, TAs reasoned differently along this spectrum depending on the context. This Chapter is presented in three sections. In Section 5.1, I provide a general overview of two different interpretations of probability – a frequency and a subjective interpretation. I discuss these interpretations with respect to two different tasks used in this study – The Unusual Sample and Gallup Poll Tasks[6], which entail investigating samples of different size and confidence intervals. In Section 5.2, I provide a conceptual analysis and framework for the Unusual Sample Task and discuss TAs' responses to the task in light of the framework. In Section 5.3, I provide a conceptual analysis and framework for the Gallup Poll Task and analyze TAs responses with this framework. The conceptual frameworks developed in this chapter are end products to this study and suggest a model of how TAs may reason and think about the

---

[6] The Unusual Sample and Gallup Poll Tasks can be found in the appendix.

connections between sampling and probability, and sampling, probability and statistical inference.

In Chapter 6, I present the final theme of this research – the implications of TAs' subject matter knowledge and knowledge of content and students on their statistical knowledge for teaching. The chapter is presented in three sections. In Section 6.1, I present a model for what necessary statistical knowledge for teaching sampling topics might look like. This model is built from the existing research literature[7] on stochastics education and teacher knowledge, as well as the findings presented in Chapters 4 and 5. In Section 6.2, I analyze the data in light of this model and discuss the types of knowledge the TAs in this study demonstrated. In Section 6.3, I briefly address the issue of TAs' beliefs. Beliefs and knowledge are intimately linked, and although my primary interest was in TAs' knowledge for teaching sampling, particular TAs' beliefs about teaching and learning came through strongly in this study.

In Chapter 7, I present the overall findings from this study and concluding remarks. Chapter 7 is presented in four sections. In Section 7.1, I discuss the overall findings and conclusions. In Section 7.2, I discuss the study's contributions and implications. In Section 7.3, I discuss the study's limitations. In Section 7.4, I discuss directions for future research on statistics TAs.

---

[7] This literature is discussed in Chapter 2.

CHAPTER 2

BACKGROUND THEORY AND LITERATURE REVIEW

The purpose of this chapter is to elaborate on the theoretical perspectives and background literature that inform and frame my research, both in design and analysis. This chapter is presented in three sections. In Section 2.1, I discuss the overarching philosophical position that guides my research – radical constructivism. This philosophical position serves to orient my beliefs about the goals and purposes of stochastics education, and it provides a model for conducting research, particularly with respect to data collection and analysis methods. In Section 2.2, I review relevant research literature on stochastics education. In particular, I discuss three constructs – *statistical thinking*, *statistical literacy* and *statistical reasoning* – in relation to sampling that frame this study on a detailed level. In Section 2.3, I address research on teacher knowledge. In general, there is little research on TAs. However, a review of research on teachers and teaching is particularly important because there are many similarities between beginning TAs and teachers. Thus, the research on teacher knowledge provides a foundation for examining TAs' knowledge.

2.1 Radical Constructivism

At the meta-level, a radical constructivist theory of learning guides this study. A central tenet of radical constructivism is that individuals construct their own knowledge of the world through their experiences. Von Glasersfeld (1990, 1995) traces the roots of radical constructivism to skeptics of the philosophical tradition of

the pursuit for absolute truth. According to von Glasersfeld (1990) skeptics of absolute

truth, such as Xenophanes (6th Century B.C.), have argued over the years that absolute

truth is

> based on the assumption that whatever ideas or knowledge we have must
> have been derived in some way from our experience, which includes
> sensing, acting, and thinking. If this is the case, we have no way of
> checking the truth of our knowledge with the world presumed to be lying
> beyond our experiential interface, because to do this, we would need an
> access to such a world that does not involve our experiencing it (1990, p.
> 20).

This view of knowledge requires letting go of the idea of absolute truth, yet it does not

necessarily lead to solipsism. Radical constructivism offers a reconstruction of

knowledge. That is, rather than view knowledge as an absolute truth that a learner

must acquire, a radical constructivist views the individual as constructing his/her own

knowledge through his/her experiences and/or interactions with the outside world.

Von Glasersfeld suggests two basic principles of radical constructivism, which he

derived from the works of Piaget (1971). The principles are,

> a. Knowledge is not passively received either through the senses or by
> way of communication. Knowledge is actively built up by the
> cognizing subject.
>
> b.
> > i. A function of cognition is adaptive, in the biological sense of
> > the term, tending toward fit or viability;
> >
> > ii. Cognition serves the subject's organization of the experiential
> > world, not the discovery of an objective ontological reality.
> > (1990, p. 23)

Of course, these principles represent a working theory and should not be construed as

absolute truth. These principles leave open the question of how knowledge is

constructed, and how individuals may work together as a community toward shared knowledge.

First, I address how knowledge is constructed for a radical constructivist. According to von Glasersfeld's (1995) interpretation of Piaget's epistemology, cognitive change takes place through the process of assimilation and accommodation. Assimilation is the process by which a particular experience is fit into an already existing conceptual structure by the person having the experience. Accommodation is the process of revising and restructuring one's existing conceptual structures. According to von Glasersfeld, learning takes place through an action scheme. The scheme consists of recognizing a particular situation, and reacting to the situation with the expectation that the reaction produces previously experienced results (p.65). For von Glasersfeld, recognizing the situation necessitates the process of assimilation, and a person will respond to the situation with a specific activity that he or she has used previously in similar situations. However, if unable to assimilate information, the person will experience perturbation, which will lead to accommodation and finally to equilibrium. For von Glasersfeld, this type of model for concept construction provides a viable model for how a person might come to know something.

The second point that needs to be addressed is that of viability. Some critics of constructivism suggest that such a model allows us to construct any reality we like. But to this von Glasersfeld (1990, 1995) argues that the construction of one's knowledge or view of reality is subject to constraints. Avoiding contradiction is one constraint that prevents us from constructing any reality we like. In addition, von

Glasersfeld (1990) argues that "[e]very individual's abstraction of experiential items is constrained (and thus guided) by social interaction and the need of collaboration and communication with other members of the group in which he or she grows up" (p. 26). Thus, as a community the goal is to find compatible versions of knowledge or explanation of phenomena, and to be flexible enough to modify these versions as new experiences suggest different explanations.

The view of knowledge and learning presented here has implications for how one thinks about statistics and what it means to engage in statistical activity (Davis, Maher, & Noddings, 1990). For example, a view of knowledge as absolute truth that exists outside of one's mind, and can be discovered through the act of learning, might lead to a belief that statistical procedures are important for securing correct answers. But to question this view of absolute truth the way that a radical constructivist would suggests alternative approaches to the study of statistics. Engaging in problem solving activities, grounded in data and context, and looking for possible explanations to problems is another way to approach statistics. Thus, my orientation towards radical constructivism impacts the way I view fundamental components of statistics, and the goals and purposes of statistics education.

Consequently, the use of constructivism as a guiding philosophy impacted my research design and analysis. This philosophy oriented me toward investigating TAs' perceptions, purposes, and ways of working out statistical and pedagogical problems, rather than simply investigating their final answers to problems, so that I could understand their thinking at a deeper level. The methodological implications of this

position are such that task-based surveys would not suffice to gather a rich characterization of TAs' statistical knowledge for teaching. Hence, I use both interview and survey data in order to build a viable model of TA reasoning and a representation of TAs' models of student reasoning. I elaborate further on my research methodologies in Chapter 3. The lens of radical constructivism also impacted my literature review in that I sought out research with a similar guiding philosophy to provide frameworks and tools for my research; this literature is addressed in the sections that follow.

2.2 Research on Stochastics Education

Research in the field of probability and statistics education has experienced an unprecedented boom over the past decade. The growth of research in this field is due in part to the growth of statistics as a discipline, NCTM's (2000) call for the inclusion of probability and statistics in school curriculum, and conferences, such as the International Conference on Teaching Statistics (ICOTS), devoted exclusively to stochastics education. In this section on stochastics education, I discuss three constructs from the literature – *statistical literacy*, *statistical thinking,* and *statistical reasoning* – which support the design of my research tasks and my analysis of the data. These constructs are examined with reference to sampling concepts and their relationship to statistical inference. These three constructs provide a framework for the type and quality of knowledge that is necessary for well-developed statistical knowledge for teaching.

2.2.1 Statistical Literacy

As societies become more information-based, technologically minded, and globally oriented, their citizens will need to have a solid understanding of basic statistics to meet society's demands and to make well-informed decisions. But what are the basic statistics required for informed citizenship? Probability and statistics educators have been addressing this question through the construct of *statistical literacy*. According to Gal (2003), the construct of statistical literacy is geared toward *consumers* of statistics, where such consumption usually takes place through the media, internet sites, newspapers, and magazines. Just as literacy is often defined as basic reading and writing skills, statistical literacy includes the basic skills necessary for understanding statistical information. Still, the term *basic* conjures up images of minimal skills; statistical literacy is in many ways much more than this. According to Ben-Zvi and Garfield (2004),

> These skills include being able to organize data, construct and display tables, and work with different representations of data. Statistical literacy also includes an understanding of concepts, vocabulary, and symbols, and includes an understanding of probability as a measure of uncertainty (p.7).

Wallman's (1993) definition, echoed by Watson and Moritz (2000a) and similar to that of Ben-Zvi and Garfield, states that, "[s]tatistical literacy is the ability to understand and critically evaluate statistical results that permeate our daily lives – coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions" (p. 1). Gal (2002) suggests that statistical literacy requires the "ability to *interpret* and *critically evaluate*

statistical information, data-related arguments, or stochastic phenomena…" (p.2).

Further, statistical literacy requires the ability to discuss or communicate one's

reactions, understandings, or opinions regarding the implications of such statistical

information (Gal, p.3). Each of these authors' definitions suggests that to be

statistically literate one must be able to *read*, *organize*, *interpret*, *critically evaluate*,

and *appreciate* statistical information presented to the public through the media. Of

course, statistical information can be represented in a variety of ways, such as

graphical and tabular forms, so statistical literacy also requires an understanding of

different representations.

One specific feature necessary for statistical literacy is informal[8] statistical

inference skills. Jacobs (1997) argues that statistical inference is essential for

effectively dealing with statistics encountered in the media. Put simply, statistical

inference is an attempt to draw conclusions about a population from data provided by

a sample (Pfannkuch, p.267, 2005). However, as I noted in the introduction, a sound

understanding of sampling concepts is necessary for an informal and formal

understanding of statistical inference. Watson and Moritz (2000a) argue that, "[a]n

understanding of sampling is fundamental to statistical literacy. Statistics are

commonly based on sample data, where sampling methods affect the quality of data

collected and subsequent inferences about populations" (p.109). Jacobs' (1997) echoes

this sentiment, "[o]ne of the main determinants of the validity of statistical inference is

---

[8] By informal I mean understanding the concept of statistical inference through an understanding of repeated sampling and the image of a distribution of sample statistics. In contrast, I take a formal understanding to mean the additional understanding of the processes and procedures for finding confidence intervals and conducting hypothesis tests.

sampling" (p. 2). Thus, an understanding of samples and sampling distributions is necessary for critical consumption of statistics in popular media.

Statistical literacy, as defined by these researchers, requires a rather sophisticated way of looking at data. This level of sophistication is important because of the expectations placed on adults living in industrialized societies. Statistical literacy serves both individuals and their communities because such knowledge supports informed public debate, improves people's ability to make decisions regarding chance-based situations, and provides an awareness of social trends, such as crime, population growth, and the spread of diseases (Gal, 2003). Thus, statistical literacy is important in the day-to-day functioning of citizens in industrialized nations, and it is my contention that statistical literacy skills should play an integral role in introductory college statistics courses.

Given that the concepts of sampling are key to developing statistical literacy and statistical inference skills, I argue that ideas of sampling should be highlighted throughout instruction. Unfortunately, most introductory statistics courses move too quickly toward formal aspects of statistical inference, and issues pertaining to statistical literacy, such as appropriate methods for "consuming statistics" found in common literature – newspapers and magazines – are not explicitly addressed as part of the introductory statistics curriculum. Specifically addressing these issues will enable learners not only to acquire statistical literacy skills, but also move beyond these skills, gaining a deeper understanding of statistical inference. As students develop informal understandings of sampling and statistical inference, and then evolve

to more formal methods of statistical inquiry, they progress toward what Gal (2003)

refers to as *producers* and *consumers* of statistics. To be both a consumer and

producer of statistics entails formal methods of statistical analysis, such as posing a

research question, designing an experiment, gathering data, using formal statistical

processes to analyze data, and drawing conclusions from the analysis. This level of

knowledge is addressed through the construct of statistical thinking.

2.2.2 Statistical Thinking

The earliest beginnings of statistical thinking were traced to John Graunt, who

studied the sources and causes of the plague in the mid 1600's (Pfannkuch & Wild,

2004). In the years following Graunt's work there was a general notion that public

policies should be informed by data, rather than the authority of the church or local

government. The emergence of statistics and statistical thinking required a shift in the

accepted thinking of the day; it required a questioning attitude and inductive thinking.

Statistical thinking emerged because of the realization that data analysis could deepen

and strengthen our knowledge of a particular situation and that probability models

could be used for modeling and predicting group behavior. As statisticians realized

that probability models could be applied to a variety of domains, and as technological

tools developed, the field of statistics matured (Pfannkuch & Wild). Today we are a

society inundated with data and it is difficult to imagine life without statistics.

Statistics is distinctly different than mathematics, and different ways of thinking

are needed in order to effectively explore and analyze data. After studying the

historical development of statistics and observing statisticians in their work,

Pfannkuch & Wild (2004) noticed five fundamental features of statistical thinking:

1. Recognition of the need for data

2. Transnumeration

3. Consideration of variation

4. Reasoning with statistical models

5. Integrating the statistical and contextual (p.18-20)

According to Pfannkuch and Wild, in the first component of statistical thinking, there

is a recognition that data need to be properly gathered because decisions cannot be

made on anecdotal evidence. An understanding of how to properly gather data is

essential to the sampling process and the attainment of valid and reliable statistics. The

second component, transnumeration, entails finding measures that reflect the

characteristics of the real situation, transforming the data into summaries and/or

graphical representations, and finding meaning in the data. Shaughnessy (2007)

elaborates on Pfannkuch and Wild's introduction of the term transnumeration,

suggesting that a different representation of the data "can reveal entirely new or

different features that were previously hidden" (p. 963). Third, consideration of

variation requires the ability to understand the causes and potential sources of

variation in the data. In addition, this third component requires an understanding of

how to act on variation – whether it should be ignored, planned for, or controlled.

Cobb and Moore (1997) also stress the importance of variation in statistics. "The focus

on variability naturally gives statistics a particular content that sets it apart from

mathematics itself and from other mathematical sciences" (Cobb & Moore, 1997,

p.801). The fourth component, reasoning with statistical models, requires the ability to

make use of aggregate data, rather than reason about individual elements. This

component also requires the ability to recognize patterns and relationships, and create

dialogue between models and data. The final component of statistical thinking requires

the ability to connect and integrate data within a particular context. Pfannkuch and

Wild's argument that context plays a central role in statistical literacy is echoed by

Cobb and Moore. They state, "[s]tatistics requires a different *kind* of thinking, because

data are not just numbers, *they are numbers with a context*" (p. 801).

Pfannkuch and Wild (2004) also noticed that statisticians cycle through the

components of statistical thinking as they engage in their work. Initially statisticians

enter the planning stage, where they investigate a problem and make plans for

collecting and analyzing data. When data is collected, the statistician must find ways

of representing or characterizing the data. She must consider variation in the process.

She must also question and critically examine the models and underlying assumptions.

In general, Pfannkuch and Wild suggest that a statistician must have curiosity,

imagination and skepticism in order to adequately model behavior and solve problems.

Pfannkuch and Wild's characterization of statistical thinking is consistent with Ben-

Zvi and Garfield's (2004) definition of statistical thinking as "an understanding of

why and how statistical investigations are conducted and the 'big ideas' that underlie

statistical investigations" (p. 7). Pfannkuch and Wild's commentary makes an

important contribution to statistics education by providing a global model of statistical

thinking. Their model characterizes the types of thinking *necessary* for statisticians and teachers of statistics. In addition, this model is a pedagogical tool. An understanding of the historical development of statistical thinking and the components necessary for successful statistical inquiry provide statistics educators with an explicit picture of the types of thinking that need to be developed by students in order for them to be both *consumers* and *producers* of statistics.

In reflecting on the constructs of statistical literacy and statistical thinking there is certainly a link between them in the sense that both entail a critical eye when examining statistical information and require an understanding of the key concepts of sampling processes. If teachers can structure classrooms in ways that promote the development of statistical thinking, then students will have the opportunity to develop questioning attitudes, construct their own understandings of statistical processes, and think like statisticians (Pfannkuch & Wild, 2004). I turn now to a discussion of the construct of statistical reasoning and what the research does suggest about students' statistical development.

2.2.3 Statistical Reasoning

According to Ben-Zvi and Garfield (2004), statistical reasoning "may be defined as the way people reason with statistical ideas and make sense of statistical information" (p.7). As statistical thinking incorporates the "big ideas" or global view of the process of statistical inquiry and a picture of the types of thinking and knowledge structures required of statisticians and statistics educators, statistical

reasoning provides a local view of how students make sense of statistical information in a particular situation. As research in statistics education has blossomed, building blocks for understanding sampling concepts have been identified and numerous cognitive models for describing types of student reasoning have emerged. Many of these models overlap and different groups of researchers identify and describe similar types of student reasoning (Shaughnessy, 2007). As the research presented here is concerned with TAs' statistical knowledge for teaching sampling concepts, it is important to examine the research literature on K-12 and undergraduate students' reasoning in this area. Insight into the necessary knowledge components for understanding ideas of sampling, common conceptual difficulties, and cognitive models all serve as tools for framing and justifying the research methodologies in Chapter 3.

Building blocks to understanding sampling concepts

A natural first step in thinking about students' conceptual development of sampling processes is to wonder what conceptual building blocks are necessary for a profound understanding of sampling. A thorough review of the literature (Chance, delMas, & Garfield, 2004; Heid, Perkinson, Peters, & Fratto, 2005; Pfannkuch, 2005; Reading & Shaughnessy, 2004; Saldanha & Thompson, 2003; Shaughnessy, 2007; Shaughnessy & Chance, 2005; Watson & Moritz, 2000b) suggests several key features necessary for developing concepts of sampling and connecting ideas of sampling to statistical inference. Figure 2 represents my synthesis of the necessary knowledge

components for a well-developed understanding of sampling as gleaned from the

stochastics education literature.

Figure 2: Knowledge Components of Sampling

Necessary Knowledge Components for Understanding Sampling Concepts and Making Connections between Sampling and Statistical Inference

1. Definition of Sample
   a. Sample as a subset of the population
   b. Sample as part of a larger collection of other samples from the same population
2. Proper Sampling Methods
   a. Random sampling
   b. Sources of bias in sampling
3. Measures of Center
   a. Mode/ Median/ Mean
4. Measures of Spread
   a. Range/ Interquartile range
   b. Variance/ absolute deviation/ standard deviation
5. **Attend to and coordinate multiple aspects of a distribution simultaneously**
   a. **Coordinating measures of center, spread and shape to reason informally or formally about a distribution**
6. Sampling Distributions
   a. Definition of sampling distribution
   b. Distinction between the distribution of a sample of observations and the distribution of sample statistics
   c. Distinction between empirical and theoretical sampling distributions
7. Properties of the Normal Distribution
   a. Shape, center and spread
   b. More than just a bell curve
8. **The role of sampling in the creation of confidence intervals**
   a. **Interpreting level of confidence through a perspective that supports the image of repeated sampling and a long-term relative frequency perspective of probability**
9. The role of confidence intervals in making statistical claims
   a. Connection between confidence intervals and hypothesis tests
10. **The role of variability**
    a. **Recognition of variability within and between samples**
       i. **Variability of sample statistics in a sampling distribution**
    b. **As sample size increases sample variability decreases – Law of Large Numbers**
    c. **Balance between variability and representativeness**
       i. **Understanding bounded variability**

The knowledge components presented here aid in the design of the research tasks (see Chapter 3). These concepts are not trivial – they require the ability to combine and integrate a multitude of statistical topics. Furthermore, the knowledge components addressed here form a basis for the development of the Central Limit Theorem, as well as the formal processes of statistical inference, such as confidence intervals and hypothesis tests – key components of introductory statistics. Notice that knowledge components 5, 8, and 10 are in bold; I highlight these components for three reasons. First, the three knowledge components highlighted in Figure 2 were emphasized by multiple researchers (Bakker & Gravemeijer, 2004; Chance, delMas, & Garfield, 2004; Makar & Confrey, 2004; Pfannkuch, 2005; Reading & Shaughnessy, 2004; Saldanha & Thompson, 2003; Shaughnessy, 2007; Shaughnessy & Chance, 2005; Watson & Moritz, 2000a&b) as core elements necessary for a profound understanding of sampling and the development of statistical literacy and statistical thinking. Second, the researchers above observed that the highlighted knowledge components shown in Figure 2 were particularly difficult concepts for students[9]. Third, the knowledge components highlighted in Figure 2 foreshadow the key role they play in this study. I elaborate further on the knowledge components presented in Figure 2 in the next section, when I discuss student reasoning, because it provides an opportunity to specify these knowledge components by grounding them in particular statistics contexts. I begin that discussion by examining what the research says about students'

---

[9] That these features are problematic for students is discussed in detail in the next subsection – Student Reasoning and Developmental Difficulties.

difficulties with sampling processes, and the challenges that knowledge components 5, 8, and 10, exhibited in Figure 2, present for students.

Student Reasoning and Developmental Difficulties

As Figure 2 represented my synthesis of the research literature on the necessary building blocks for understanding sampling and statistical inference problems, Figure 3 represents my synthesis of the literature (Bakker & Gravemeijer, 2004; Chance, delMas, & Garfield, 2004; Pfannkuch, 2005; Reading & Shaughnessy, 2000, 2004; Rubin, Bruce and Tenney (1991); Saldanha & Thompson, 2003; Shaughnessy, 2007; Shaughnessy & Chance, 2005; Watson & Moritz, 2000a&b) pertaining to the types of difficulties students are likely to experience as they develop their understandings of sampling and statistical inference. Notice that for each conceptual building block, there appears to be a corresponding conceptual hurdle for students[10].

---

[10] I specify more carefully the common conceptual hurdles in the following subsections sections on student reasoning, by grounding them in the context of different sampling tasks.

Figure 3: Common Difficulties in Students' Development of Sampling

Common Difficulties and Misconceptions in Students' Reasoning and Development of Sampling Concepts

- Difficulty with the concept of random sample
    - Difficulty distinguishing between the colloquial versus statistical use of the term random and the term sample
    - Difficulty recognizing sources of bias in sampling
- Difficulty with the added level of abstraction required for understanding sampling distributions
    - Difficulty with the difference between a distribution of a sample and the distribution of a collection of sample statistics
    - Difficulty with the distinction between empirical and theoretical sampling distributions (a sophisticated concept – difficulty documented in teachers (Heid et al., 2005))
- **Difficulty attending to multiple aspects of a distribution**
    - Overly focused on modes or other measures of center
    - Overly focused on variability or individual data points
    - Focus on shape – Difficulty making distinctions between the normal and other symmetric shaped distributions
- **Difficulty finding a balance between sample representativeness and sample variability**
    - **Difficulty understanding the role of sample size in sampling variability**
        - **Do not expect a difference in variability for different size samples or believe that large samples have more variability**
    - **Difficulty relating a long-term relative frequency view of probability to sampling and statistical inference problems**
- **Difficulty understanding the role of sampling in the creation confidence intervals**
    - **Difficulty conceptualizing confidence level and margin of error – maintaining an image of repeating the sampling process**
- Difficulty with the concept of the Central Limit Theorem

I highlight in bold students' difficulties with: (a) the role of variability; (b) the role of sampling in interpreting confidence intervals; and, (c) coordinating multiple

attributes of a distribution simultaneously. These particular difficulties are highlighted for three primary reasons. First, as I mentioned at the end of the previous section multiple researchers have observed that these components are particularly difficult for students. Second, Liu and Thompson (2005), Makar and Confrey (2004), and Heid, Perkinson, Peters and Fratto (2005) observed that the role of sampling and sampling distributions and their relationship to confidence intervals were difficult concepts for secondary teachers. Third, these features also prove to be difficult and challenging for TAs[11]. In the sections that follow, I elaborate in more detail about the types of difficulties students and/or teachers have with the role of variability, the role of sampling in the creation of confidence intervals, and coordinating multiple attributes of a distribution.

The Role of Variability

In reviewing the work of Chance, delMas, and Garfield (2004), Pfannkuch (2005), Reading and Shaughnessy (2000, 2004), Rubin, Bruce and Tenney (1991), Shaugnessy (2007), Saldanha and Thompson (2003), and Watson and Moritz (2000a&b), it is clear that variation is a major component for understanding sampling processes. Recall that this characteristic is also one of the five components of statistical thinking (Pfannkuch & Wild, 2004) and a necessary component for the development of statistical literacy (Watson & Moritz 2000a, Jacobs, 1997). It also is a rather difficult and elusive concept for students to grasp.

---

[11] This paper will address this third point and provide evidence in the analysis section of the types of difficulties TAs experienced with these conceptual building blocks.

Before discussing student conceptions of variation I briefly discuss issues of variability in sampling and the relationship between sample size and sample variability. In probability theory, the Law of Large Numbers states that "the relative frequency of the number of times that an outcome occurs when an experiment is replicated over and over again (i.e., a large number of times) approaches the true (or theoretical) probability of the outcome" (McClave & Sincich, 2000, p.102). Figure 4 is a representation of this relative frequency perspective.

Figure 4: Visual Representation of the Law of Large Numbers

Relative
Frequency

small n
large n
n

n = Sample Size
As n increases the relative frequency stabilizes
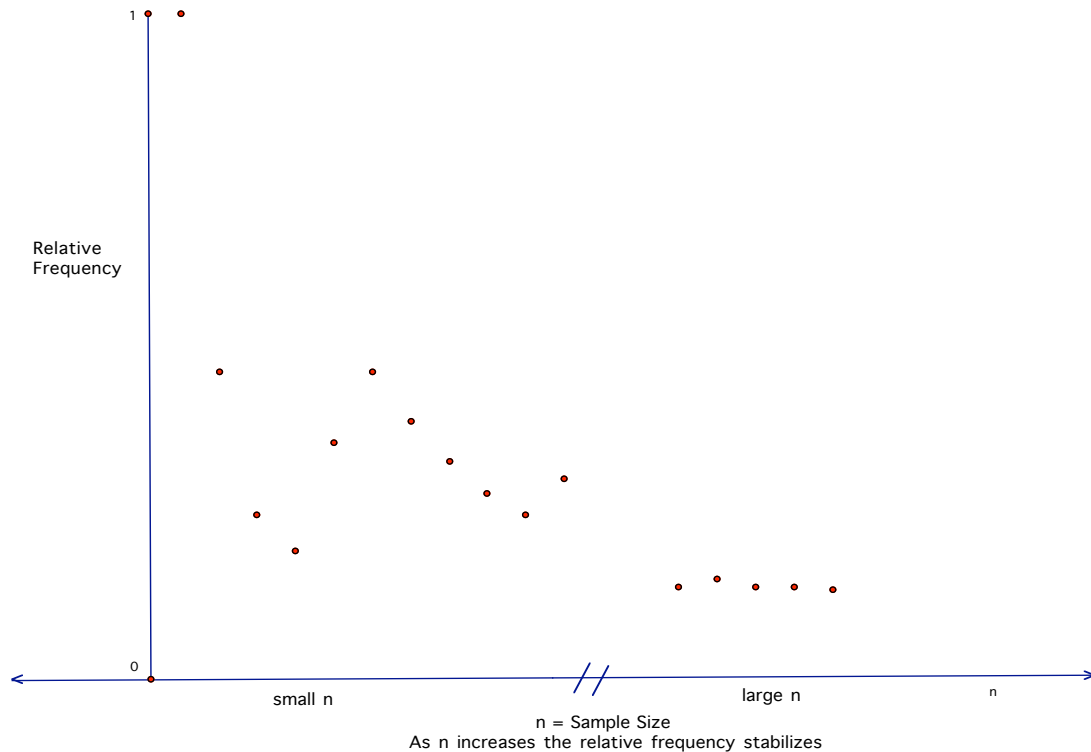
In a sampling context, the Law of Large Numbers implies that the larger the sample, the more likely it is that the sample is *representative* of the population in its

characteristics, such as mean and standard deviation (assuming the sample was chosen randomly). To illustrate the relationship between variability and sample size, consider a jar containing 10000 candies, 7500 red and 2500 yellow. Now consider a sample of 10 candies versus a sample of 1000 candies picked at random from the jar. By the Law of Large Numbers, the larger sample of 1000 candies is *more likely* to approximate the true ratio of red to yellow candies in the container. In addition, an unusual ratio, say 2 reds to 8 yellows, is *more likely* to come from the smaller sample of 10 candies. Another way to make sense of variability in this context is to imagine the distribution of all possible sample statistics taken from samples of size *n*. As *n* increases, the standard error of the sampling distribution decreases, resulting in a greater concentration of sample statistics in a closer proximity to the population mean.

A finding of major consequence in the research literature is that students' conceptions of sampling fall within a spectrum, which posits representativeness at one end and variability at the other (Jacobs, 1997; Kahneman & Tversky, 1972; Konold, 1989, 1991; Rubin, Bruce, & Tenney, 1991; Reading & Shaughnessy, 2000 & 2004; Shaughnessy, Watson, Moritz, & Reading, 1999; Watson & Kelly, 2004). As students traverse this continuum, they must grapple with the role that sample size and sample selection methods play in the variability of sampling processes. Unfortunately, much of the research indicates that many students reason at the extremes of this spectrum and not in the middle. That is, students experience a difficult time with the concept of bounded variability (in the sense of Saldanha & Thompson, 2003; Thompson, Liu, & Saldanha, 2007).

On the one hand, students have a tendency to believe that there is more variability in sampling situations than is probable. Jacobs' (1997) studied the ways in which Grade 4 and 5 students evaluate sampling methods in the context of surveys. She observed that children had difficulty with the concept of simple random sampling because they were overly focused on the possibility of obtaining an unusual sample. For example, the children in her study expressed concern that a simple random sample of grade 4 children would be unfair because the sample might end up consisting of all girls. The children's concern that the sample would be unfair, due to the chance of getting all girls or all boys, indicates that they were overly focused on extreme values, thinking that such samples were more likely to happen than probability suggests.

In their study of 12th grade students (N = 12), Rubin, Bruce, and Tenney (1991) likewise observed the tendency for students to expect too much variability in sampling situations, and to focus too heavily on rare occurrences. For example, Rubin, Bruce, and Tenney asked students the following question,

> Four hundred campers are to be divided into two teams, red and blue, for a track meet. One counselor says that the campers should be divided between the two teams using the following method: You put all the campers' names into a hat and mix them up real well. Then you pick one name at a time out of the hat. The first name picked goes to the red team, the second to the blue, the third to the red, and so on. Another counselor argues that the campers should be divided according to how fast they can run so that the teams have about the same number of fast and slow runners. They argue back and forth.
>
> Show me how many fast and slow runners you think would get on each team using the two methods. You can assume there are 200 fast and 200 slow runners (p. 17).

Most of the students in their study claimed that picking names out of a hat could produce fair teams, yet most of the students also believed that teams with 150 fast runners and 50 slow runners would be quite likely. The majority of the responses that students gave as "likely" for the break down of the two teams were in fact extremely unlikely to occur. According to Rubin, Bruce and Tenney the majority of the "likely" responses given by students would occur less frequently than 1 in 100 samples.

On the other hand, students have a tendency to believe that samples are identical to the parent population. Rubin, Bruce, and Tenney (1991) also observed that many students believe a sample provides all the information one needs about a population because they fail to think about issues of variability. To illustrate, Rubin, Bruce, and Tenney posed the following Gummy Bear problem to the same group of 12th grade students (N = 12),

> Suppose you took your little brother or sister to an Easter parade in Boston. At the parade, the "Easter Bunny" handed out packets of Gummy Bears to all the kids. Each packet had 6 Gummy Bears in it. To make up the packets, the Easter Bunny took 2 million green Gummy Bears and 1 million red Gummy Bears, put them in a very big barrel and mixed them up from night until morning. Then he spent the next few hours making up the packets of six Gummy Bears. He did this by grabbing a handful of Gummy Bears and filling as many packets as he could. Then he reached into the barrel and took another handful, and so on, until all the packets were filled with 6 Gummy Bears. When you get home from the parade, you open up your packet.
>
> 1. How many green Gummy Bears do you think might be in your packet? Can you tell me how you got that?
> 2. Do you think all the kids got that many greens? Can you explain that to me?
> 3. If you could look at the packets of 100 kids, how many kids do you think got n (the number from above) greens?

4. How many different color combinations of red and green Gummy Bears are there?
5. Out of 100 kids, how many do you think got 0 green? 1 green? 2 green? 3 green? 4 green? 5 green? 6 green? (p.16)

All of the students interviewed recognized the ratio of red to green Gummy Bears. Using the ratio as their reason, all of the students answered *four* to the first question. While all of the students also recognized that four green Gummy Bears would not be present in *every* packet, their reasons were different. Students tended to evoke failures in the Easter Bunny's sampling procedure to explain the variability, rather than raising the concept of random error. For example, some students mentioned that Gummy Bears are sticky so a bunch of green might all get stuck together. What is particularly telling are the estimates students gave for the number of kids they suspected would have four green Gummy Bears in their packets. Students tended to think that more than 75 out of 100 kids would have *exactly* four green Gummy Bears in their packets. Figure 5 shows the theoretical binomial probability distribution for the Gummy Bear problem, along with an example of a typical student response for the number of packets containing 0, 1, 2, 3, 4, 5, and 6 green Gummy Bears. Notice that in this problem students had a tendency to be overly focused on center.

Figure 5: Gummy Bear Sampling Distributions

(Reproduced from: Rubin, Bruce, & Tenney, 1991, p. 22)
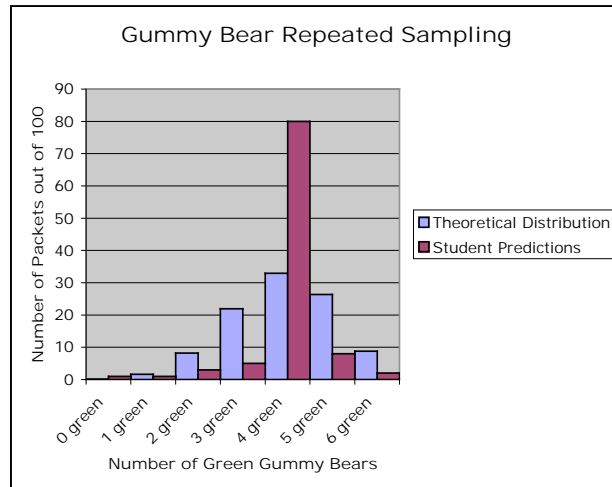
In a comparable sampling task, Shaughnessy, Ciancetta, and Canada (2004a)

observed similar results in 6-12th grade students (N = 272). The task reads as follows:

> Suppose you have a container with 100 candies in it. 60 are red, and 40 are
> yellow. The candies are all mixed up in the container. You pull out a
> handful of 10 candies. Suppose that 50 students each pulled out 10 candies,
> from the bowl, wrote down the number of reds, put them back, mixed them
> up. Of the 50 students, how many of them do you think would get:

      0 reds and 10 yellows?_____
      1 reds and 9 yellows?_____
      2 reds and 8 yellows?_____
      3 reds and 7 yellows?_____
      4 reds and 6 yellows?_____
      5 reds and 5 yellows?_____
      6 reds and 4 yellows?_____
      7 reds and 3 yellows?_____
      8 reds and 2 yellows?_____
      9 reds and 1 yellow?_____
     10 reds and 0 yellows?_____

       Total   50

In this task, there was a tendency for students to either stack up their responses in the center with narrow distributions, or to create wide distributions, expecting at least one handful for each possibility – 0 through 10 red candies. When students created 'wide' distributions they either (a) created a fairly uniform distribution by evenly distributing across all outcomes, or (b) more often, students stacked most of the outcomes at the center and placed just a few outcomes in the extreme locations. Figure 6 shows the distribution for a 'narrow' response and Figure 7 shows the distribution for a 'wide' response that is stacked on the center.

Figure 6: Candy Mixture – Narrow Center Focus
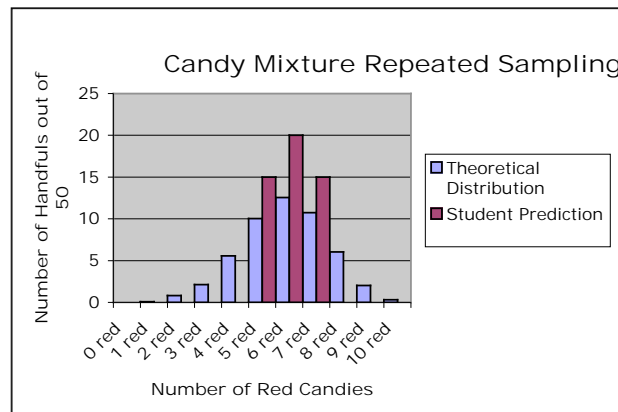
Figure 7: Candy Mixture – Wide Center Focus
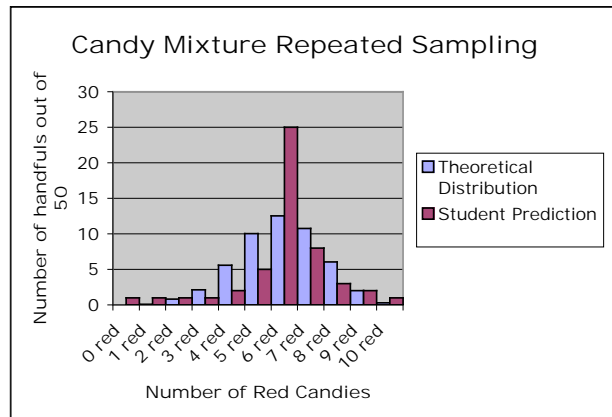
The two examples presented here reflect the tension that students experience between representativeness and variability in the sampling process, and their tendency to position themselves at one extreme or the other on a spectrum of representativeness to variability. This suggests that the concept of bounded variability in sampling is not trivial.

Kahneman and Tversky (1971, 1972), and Watson and Moritz (2000a&b) also observed students' belief that a sample, however small or poorly collected, is completely representative of the population from which it is drawn. For example, Kahneman and Tversky (1972) asked college students to consider two different sized hospitals and make a prediction about the number of baby boys born each day at the hospitals over the course of a year. The problem is shown below.

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower.

> For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?
>   a. The larger hospital
>   b. The smaller hospital
>   c. About the same (that is, within 5 percent of each other) (p. 443).

Of the 95 undergraduate students in Kahneman and Tversky's study, 21 picked choice (a), 21 picked choice (b), and 53 picked choice (c). Choice (c), picked by the majority of students, serves as an illustration of what Kahneman and Tversky termed the *representativeness heuristic*, because it suggests that students believe the number of boys born at both the large and small hospitals is equally representative of the general population (that is, an equal number of boys and girls). In actuality, one should expect the smaller hospital, with fewer babies born each day, to have more variability from day to day.

Watson and Moritz (2000a&b) used a comparable task to that of Kahneman and Tversky's (1972) with students in grades 3-11. In Watson and Moritz's (2000b) problem, students were told that researchers were studying the weights of grade 5 children.

> Researchers went to two schools, one school in the center of the city and one school in the country. Each school had about half girls and half boys. The researchers took a random sample of 50 children from the city school, and 20 children from the country school. One of these samples was unusual: it had more than 80% boys. Is it more likely to have come from:
>   1. The large sample of 50 from the city school, or
>   2. The small sample of 20 from the country school, or
>   3. Are both samples equally likely to have been the unusual sample? (p. 52)

Watson and Moritz achieved results similar to those of Kahneman and Tversky. That is, the most common response from students was that both samples are equally likely to be the unusual one. Student explanations indicated that if the sampling process is random, then there is no reason to suspect an unusual sample from either group. Students' choices in these tasks suggest that they fail to recognize the role of sample size in sample variability.

Konold and his colleagues (Konold, 1989; Konold, 1991; Konold, Pollatsek, Well, Lohmeier, & Lipson, 1993) found compelling evidence that supports the findings of Kahneman and Tversky (1972), and Watson and Moritz (2000b) with the added benefit of explaining why students might be inclined to think that the sample size does not affect sample variability. Konold (1989, 1991) suggested that students might reason about situations of uncertainty in a non-probabilistic manner, thinking that they must "successfully *predict* the outcome of a *single* trial" (1991, p. 146), rather than thinking about long-term relative frequency and/or a distribution of sample statistics. Students that interpret a sampling problem as pertaining to the particular sample, rather than an image of repeated samples, are likely to reason that it is not impossible for either the large or small sample to be unusual, and therefore there is no way to tell which of the two individual samples *will* be unusual. Konold referred to this form of reasoning as the "outcome approach". Gigerenzer's (1991, 1996) work supports the argument made by Konold. Gigerenzer argued that if a student responded that both hospitals are equally likely to have an unusually high number of baby boys born, and the student was making that decision based on his understanding that he was to predict

which of those *individual* samples *was* going to be unusual, then this response would be legitimate. That is, Gigerenzer argued that a subjective view of probability, rather than a frequency perspective, might be a more natural inclination for individuals as they reason about uncertainty in particular contexts.

In sum, the research of Jacobs (1997), Rubin, Bruce and Tenney (1991), Kahneman and Tversky (1972), Shaughnessy et al. (2004a), and Watson and Moritz (2000a&b) strongly suggest that students tend to fall at the extremes of this variability – representativeness spectrum. That is, on the one hand, some students become overly focused on variability and the occurrence of extreme values in the sampling process, concluding that little can be inferred from a sample. On the other hand, some students become overly focused on representativeness, and as a result they fail to see instances of bias in sampling due to poor data collection methods or variability due to random error. Rubin, Bruce, and Tenney noticed that the *same* student would be overly focused on centers in one problem and then variability in another. This suggests that context plays an important role in the way students approach a problem. In some instances, students might feel as though they are being asked to predict with certainty the outcome of a single event and in other instances they might be inclined to think about a distribution of possible outcomes. Konold (1989, 1991) and Gigerenzer (1991, 1996) also observed students' inclinations toward interpreting probabilities from a subjective, non-probabilistic perspective in certain contexts.

Also worth noting is that students' difficulty resolving tensions between sample representativeness and sample variability span grade levels, from elementary to

college students. It would seem that tensions experienced by students between one extreme and the other needs to be addressed in the classroom in order for students to develop some intuition for both the role of expectation and variability. Shaughnessy (2007) notes that if students have opportunities to experiment with sampling activities, they are often able to resolve these tensions so that they can make statistically educated decisions when confronted with information from a sample.

The Role of Sampling

The developmental transition from samples to sampling distributions proves to be a rather abstract and difficult concept for students (Batanero, Tauber, & Sanchez, 2004; Chance, delMas, & Garfield, 2004; Shaughnessy & Chance, 2005). Sampling distributions remain a nebulous concept for most students because it "requires students to combine earlier course topics such as sample, population, distribution, variability, and sampling" (Chance, delMas, & Garfield, 2004, p.295), which many students may have a tentative understanding of at best. Shaughnessy and Chance noted students' confusion between the concepts of samples and sampling distributions, and between empirical and theoretical sampling distributions. Shaughnessy and Chance suggest that students may experience confusion over the different terms or by the extended level of abstraction in understanding that the mean of a sampling distribution is the mean of a collection of means. In addition, in their work with secondary teachers, Heid et al. (2005) observed that the distinction between empirical and theoretical sampling distributions was difficult for teachers to grasp.

Saldanha and Thompson (2003) distinguish two ways in which students might conceive of samples and sampling. The first image of sample "entails images of repeating the sampling process and an image of variability among its outcomes" (p.257). The second image simply views a sample as a subset of a population. Saldanha and Thompson argue that the former image supports reasoning about distributions and the development of statistical inference because it relates the particular sample at hand to the larger picture of all possible samples, viewing it as one particular case from a group of similar cases. Saldanha and Thompson argue that the latter conception of sample takes a singular approach to the problem and students with such an image tend to focus on individual samples, believing they must predict the outcome of an event with absolute certainty. That is, this singular approach to sampling problems is likely to lead to an interpretation consistent with Konold's (1989) "outcome approach". Saldanha and Thompson suggest that viewing the sample as one particular case from the larger group of sample statistics lays the foundation for the examination of a distribution of sample statistics, and for using that distribution as the foundation for statistical inference claims.

Despite students' difficulties reasoning about sampling concepts, and the importance placed on sampling concepts by the statistics education community, Watson and Moritz (1997) found that the teachers they studied did not see a need to introduce sampling in their lessons. In particular, introducing sampling in connection with statistical inference was non-existent among the teachers. In addition, Liu and Thompson (2005), and Makar and Confrey (2004) found that secondary teachers'

knowledge of sampling in connection to statistical inference was tenuous. Makar and

Confrey indicated that the teachers in their study experienced difficulty comparing

sampling distributions and making informal statistical inference based on their

reasoning about the properties of the different distributions. Makar and Confrey

suggested that this difficulty was in part due to teachers experiencing problems with

thinking about variability within and between sampling distributions.

Liu and Thompson (2005) investigated teachers' informal and formal

understandings of statistical inference. Specifically, they investigated teachers'

knowledge of confidence intervals, including confidence level and margin of error.

Liu and Thompson used the following task in assessing teachers' understandings:

> Stan's statistics class was discussing a Gallup poll of 500 TN voters'
> opinions regarding the creation of state income tax. The poll stated, "…the
> survey showed that 36% of Tennessee voters think a state income tax is
> necessary to overcome future budget problems. The poll had a margin of
> error of ± 4%."

> Stan said that the margin of error being 4% means that between 32% and
> 40% of TN voters believe an income tax is necessary. Is Stan's
> interpretation a good one? If so, explain. If not, what should it be? (p. 4)

While none of the teachers in Liu's and Thompson's study believed that Stan's

interpretation was completely correct, all teachers initially gave incorrect alternative

interpretations. A correct interpretation would be: if we took random samples of size

500 and we repeated this process 100 times, then approximately 93 of those times the

sample proportion ± 4% would capture the true population proportion. Unfortunately,

Liu found that some teachers interpreted the problem as $x\%$ of sample proportions

would fall within 32% to 40%. Other teachers did not understand that they could find the confidence level using a statistical table for the standard normal distribution, or they just assumed the confidence level to be 95%. Finally, other teachers expressed the idea that the population proportion changed from sample to sample, rather than the sample proportion. Liu argued that confidence intervals and margin of error requires an understanding of sampling and an image of repeated sampling, which these teachers lacked.

As universities become responsible for quality undergraduate education, professors and TAs will be accountable for their teaching. This study begins a research base on tertiary teachers' conceptions of statistics. As it is likely that there are similarities between secondary teachers' and TAs' knowledge of statistical concepts, it is important to understand how teachers reason about sampling as a basis for studying TAs. The literature presented here provides several potential tasks, and features of the sampling process and its relation to statistical inference that would be both interesting and significant to address with TAs. In addition, frameworks built on research about students' and secondary teachers' reasoning provides a foundational structure from which to analyze TA responses to similar sampling tasks.

Coordinating Multiple Aspects of a Distribution

The ability to reason distributionally is stressed by statistics education researchers (Bakker & Gravemeijer, 2004; Shaughnessy, 2007; Shaughnessy et al., 2005) as an

important building block[12] for understanding sampling concepts. Yet, this same research shows that students struggle to coordinate multiple attributes of a distribution as a means for answering statistical questions. Students have a tendency to focus on individual data points or the mode of a data set (Bakker & Gravemeijer, 2004; Reading & Shaughnessy, 2004).

Shaughnessy et al. (2004a&b, 2005) noted three common features in the development of students' statistical reasoning – *additive*, *proportional*, and *distributional* and developed a framework around their observations. Additive reasoners attend to individual frequencies of the sample or sampling distribution. For example, students who reason additively tend to focus on the mode, as it is the most frequent value. Proportional reasoners primarily use the underlying ratios of the population as they reason in sampling situations. Distributional reasoners focus on two or more aspects of the sample or sampling distribution. For example, a distributional reasoner might attend to both the population proportion and the shape, or the population proportion and the variability, as they reason in a sampling context. A visual model of the coding framework developed by Shaughnessy et al. is shown in Figure 8.

---

[12] Recall knowledge component 5 in Figure 2.

Figure 8: Types of Student Reasoning in the Context of Sampling Distributions

Conceptual Framework of Shaughnessy et al. (2004, 2005)

Other (0)

Additive (1)

Transitional Stage

Shape (2)   Mode (2)   Variation (2)

Median/Mean (3)
Proportional
(Explicit connection between the sample and population %)

Distributional (4)
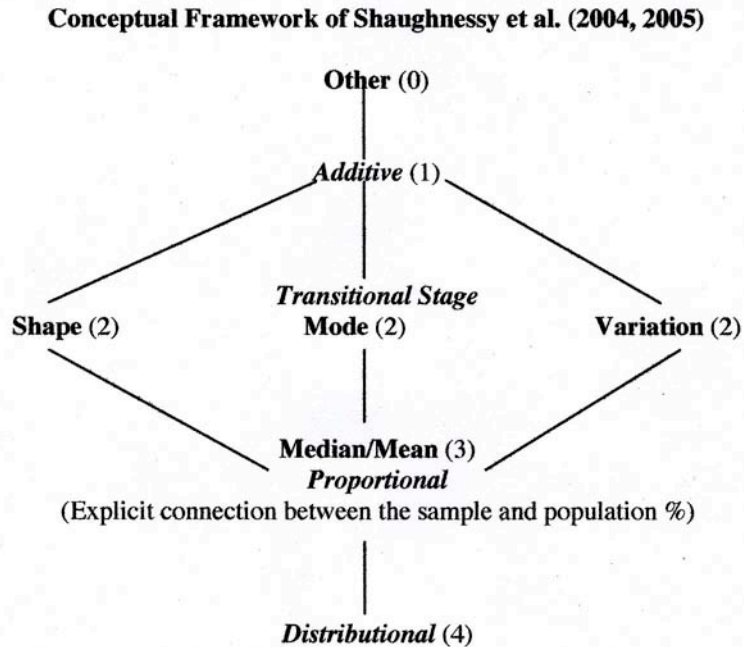
The coding framework used by Shaughnessy et al. (2004 a & b, 2005) also contained a *transitional* reasoning category for students who exhibited a higher level of reasoning than additive, but were still unable to attend to underlying population proportions. For example, a student who focused on the range or the shape of a distribution, yet made no explicit mention of the underlying population proportion, indicated that the student was able to focus on more than just individual elements of the data. The *Other* category was used for idiosyncratic and off-task responses. Each stage of reasoning corresponds to a numeric code so that students' responses could be scored from 0 to 4. These codes are hierarchal and a higher code indicates a more sophisticated response.

To illustrate the features of this conceptual framework, consider the Mystery Mixture Task (see Figure 9) given to 6-12 grade students in the Shaughnessy et al. (2004a&b, 2005) study.

Figure 9: Mystery Mixture Task

Mystery Mixture Task
The graphs below all came from a class that is trying to estimate a mystery mixture of 1000 red and yellow candies in a large jar. They pulled 50 samples of size 10 (recording the number of reds and then replacing and remixing each time). Here are the graphs for the number of reds for four groups from that class.

a) What do you think the mixture in the jar might be?
b) Explain why you think this.

Student responses on this task ranged from additive and proportional responses, to distributional responses. For example, several additive reasoners picked either 200 or

300 reds because these represented the two modal values among the four graphs.

Proportional reasoners tended to look for the median or mean value in each of the

graphs and then average them. Finally, distributional reasoners tended to look for the

median or mean for each of the graphs, but coupled that response with attention to the

overall shape or spread of the graphs. These students either provided a point estimate,

based on their overall sense of the center and range of the graphs, or, more promising,

an interval estimate for the mixture, using the idea of a confidence interval informally

as a means for capturing the population parameter, while expressing a sense for the

variability in the data set.

Other researchers (Bakker & Gravemeijer 2004; Reading and Shaughnessy, 2004;

Saldanha & Thompson, 2003) have characterized types of student reasoning in a

similar manner to the additive, proportional, and distributional reasoning constructs

used by Shaughnessy et al. (2004a&b, 2005). It seems that these are useful constructs

for characterizing student reasoning; thus, I draw on this framework as a beginning

data analysis tool in investigating TAs' statistical knowledge for teaching.

2.2.4 Summary

The construct of statistical literacy is useful for characterizing the type of

statistical knowledge that all educated adults should have. The construct of statistical

thinking is useful for characterizing the type of knowledge needed by statisticians and

statistics educators. The construct of statistical reasoning is useful for understanding

the ways in which students' reason and potential sources of difficulty in the

development of statistical literacy and thinking skills. These constructs are also interrelated in that statistical literacy and statistical thinking both require, at least on an informal level, the ability to: (1) reason proportionally; (2) coordinate multiple attributes of a distribution; (3) compare and contrast sampling distributions; and, (4) use features (1) – (3) to draw acceptable inferences. Furthermore, students should be able to apply these features of knowledge with experimental data and maintain a serious consideration of bounded variation. Pfannkuch (2005) suggests that these four features, along with a consideration of variation, play a fundamental role in the teaching of probability and statistics, and thus, these features are also connected to statistical reasoning. These features must be developed by students in order to facilitate their evolution into statistically literate adults, and, in some cases, statisticians, if they continue to learn more formal statistical processes.

The research presented in this section indicates that students, and in some cases teachers, have a difficult time with certain aspects of these knowledge components (Recall Figures 2 and 3). Specifically, coordinating multiple attributes of a distribution, conceptualizing sampling distributions, making inferences about different populations, and maintaining a sense of bounded variability throughout, present challenges in developing students' statistical literacy and thinking skills. If these statistical building blocks are problematic for teachers, then it is less likely that teachers will be successful at developing these conceptions in their students. Further, if teachers are not aware of the necessary building blocks for understanding sampling and the types of developmental stages in students' reasoning, then they are less likely

to highlight salient features of statistical thinking in their lessons, putting students at a disadvantage for developing robust statistical literacy and/or thinking skills. It is natural to wonder about TAs' knowledge of these features both in terms of their own understanding of statistics and their knowledge of student reasoning.

Thus, the research presented here serves to frame my investigation into TAs' statistical knowledge for teaching by: (1) providing a model of necessary knowledge components for teaching statistics; (2) providing tasks for measuring students' and teachers' knowledge of sampling concepts and the role of sampling in statistical inference; and, (3) providing initial frameworks for analyzing TAs' reasoning. In particular, the model I suggest for TAs' statistical knowledge for teaching requires both statistical literacy and statistical thinking expertise, which necessarily includes the knowledge components discussed in Figure 1 of this section. I turn now to more general research on teacher knowledge.

2.3 Research on Teacher Knowledge

In this section I examine teacher knowledge relevant to the present study. As this study is concerned with TAs, it was natural to look at existing literature on TAs. However, a few notable exceptions are Belnap (2005), DeFranco and McGivney-Burelle (2001), and Speer (2001). Belnap (2005), and DeFranco and McGivney-Burelle (2001) studied TAs' beliefs about teaching and learning more generally. Belnap studied the impact of training and professional development on TAs' beliefs and their teaching experiences. DeFranco and McGivney-Burelle's examined twenty-two TAs who participated in five seminar classes. These classes addressed issues of

pedagogy, curriculum, assessment, and epistemology. Although DeFranco and McGivney-Burelle found that TAs' beliefs about the teaching and learning of mathematics changed significantly by the end of the seminar, they were not able to put new beliefs into action in ways that would inform and change their teaching practices. Speer's (2001) research focused on TAs' beliefs and perspectives within the specific context of teaching reform calculus. Speer was interested in the relationships between TAs' beliefs and their moment-to-moment decisions in class. Speer argued that professional development more narrowly focused on a particular content domain might have a more substantial impact on TAs' beliefs and practices. In fact, Speer (2001) argued that the success of Cognitively Guided Instruction[13] (CGI) was in part due to its extended focus on very specific content area and the understanding that people make sense of new information through their existing knowledge, beliefs, and practices.

The brevity of research on TAs suggests that this is a ripe area open to investigating and in need of a foundation. To build an infrastructure for research on TAs' statistical knowledge for teaching requires an investigation of the literature pertaining to mathematics teachers' knowledge because there are likely to be similarities between mathematics teachers, particularly secondary teachers, and TAs. In fact, Belnap (2005) suggested that TAs pass through many of the same developmental stages in learning to teach as do novice K-12 teachers.

---

[13] Cognitively Guided Instruction is a research program that investigated teachers' pedagogical content knowledge and beliefs of addition and subtraction through a professional development program. It will be discussed in detail later in the section.

In this section, I examine research on teacher knowledge that is relevant to the present study. I discuss a brief history of research on teaching and progress to research on teacher knowledge relevant to the current study. I discuss the constructs of pedagogical content knowledge (PCK) and mathematical knowledge for teaching (MKT), as well as research that employs these constructs. In addition, I discuss the constructs of procedural and conceptual knowledge. The PCK, MKT, and procedural and conceptual constructs are useful in the present study, so I conclude this section with a discussion of how these constructs meld with the research base discussed in the prior section on stochastics education.

2.3.1 History of Research on Teachers and Teaching

Prior to the 1980's, a process-product paradigm represented the educational research philosophy of the day. In this model the process is considered effective teaching behaviors and the product is considered student achievement. Researchers were interested in teaching behaviors that produced gains in student achievement (Koehler & Grouws, 1992). For instance, studies examined teaching behaviors, such as the amount of time devoted to practice activities, or the frequency of manipulatives used during class (Koehler & Grouws). Begle's (1979) work is another example of this type of process-product research and a frequently cited study in mathematics education literature. Begle studied the number of mathematics courses teachers had as a proxy for student achievement. Based on these types of studies, researchers designed instructional materials to aid teachers as they identified particular teacher behaviors

59

that correlated with student achievement (Koehler & Grouws, 1992; Shavelson, Webb, & Burstein, 1986).

In the 1980's, researchers began asking different types of research questions and recognized that the process-product paradigm was insufficient for constructing more meaningful understandings of the complex act of teaching. The process-product paradigm was criticized for its limited view of the teaching process and role of the teacher (Shavelson, Webb, & Burstein 1986; Shulman, 1986). Research in the area of teaching evolved in the 1980's and 1990's to include more qualitative research on teachers' thought processes - teachers' theories, beliefs, knowledge, and decision-making processes (Clark & Peterson, 1986).

2.3.2 Pedagogical Content Knowledge and Mathematical Knowledge for Teaching

One significant change in research on teacher knowledge was Shulman's (1986) *pedagogical content knowledge* (PCK) paradigm. Shulman described how past research on teacher knowledge either focused on teachers' specific content knowledge or their pedagogical knowledge. That is, research either focused on the number of mathematics courses a teacher had and/or their mathematics test scores; or, research focused on teachers' classroom management and/or their organization and presentation of instructional plans as an indication of teachers' abilities to teach mathematics. Shulman's PCK provided a link between content and pedagogy. Shulman provided a compelling argument that the expert knowledge of a

mathematician is not sufficient for teaching mathematics, and that qualities such as classroom management, which is completely void of subject matter, would be insufficient for a thorough understanding of the knowledge required to teach mathematics. In particular, Shulman defined PCK as content knowledge that "goes beyond knowledge of the subject matter per se to the dimension of subject matter knowledge *for teaching*" (p. 9). That is, for Shulman, PCK was a type of content knowledge that embodied the "teachability" of the material to students that are encountering the concepts for the first time.

Shulman's (1986) concept of PCK significantly impacted the work of Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Hill, Rowan & Ball, 2005) and their introduction of the construct *mathematical knowledge for teaching* (MKT). Shulman's PCK and Ball and her colleagues' MKT are similar in that both constructs highlight the special knowledge mathematics teachers need in order to successfully do their jobs; however, Ball and her colleagues expand and refine Shulman's original work in at least three ways. First, the past decade of work by Ball and her colleagues represents a significant and original contribution to the development of a characterization of the mathematical knowledge necessary *specifically* for teaching elementary mathematics curriculum. Second, Ball and her colleagues argue that research on teachers' knowledge must be situated in practice and grounded in research on student thinking in order to significantly contribute toward a shared understanding of the mathematical knowledge necessary and sufficient for successful teaching. Third, Ball and her colleagues deconstruct MKT

into four primary components: (1) common content knowledge, (2) specialized content knowledge, (3) knowledge of content and students, and (4) knowledge of content and teaching.

Ball (2005) defines common content knowledge as "the mathematical knowledge and skill expected of any well-educated adult", including the ability to "recognize wrong answers, spot inaccurate definitions in textbooks, use notation correctly and the ability to do the work assigned to students" (p.13). Ball defines specialized content knowledge as "the mathematical knowledge and skill needed by teachers in their work and beyond that expected of any well-educated adult" including the ability to "analyze errors and evaluate alternative ideas, give mathematical explanations and use mathematical representations, and be explicit about mathematical language and practices" (p. 14). Ball defines knowledge of content and students, and knowledge of content and teaching, as knowledge that combines knowledge of content and students or content and teaching, respectively (pp.16-18). This combination of knowledge includes the ability to "interpret students' incomplete thinking" (p. 16) and "sequence content for instruction" (p. 18). The components of mathematical knowledge for teaching developed by Ball and her colleague's (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Hill, Rowan & Ball, 2005) contribute to my framework of statistical knowledge for teaching by illuminating foundational components necessary for teaching any subject – content knowledge and knowledge of content and students. Also, their model provides a methodological consideration for the current study to consider – research design that enables the investigation of TAs'

knowledge of content and students. These contributions are further specified at the end of Section 2.3.

The Cognitively Guided Instruction (CGI) research (Carpenter, Fennema, Peterson, & Carey, 1988; Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; Peterson, Fennema, Carpenter, & Loef, 1989; Fennema, Carpenter, Franke, Levi, Jacobs, & Empson, 1996) in the 1980's and 1990's, and the work of Even (1993), are illustrations of the shift toward characterizing the special type of knowledge needed by teachers in their work. The CGI studies focused on whether or not teachers' knowledge of research on student thinking in a particular content domain could form the basis for classroom instruction, and if such instruction would yield positive gains in student achievement (Fennema & Franke, 1992; Fennema, Peterson, Chiang, Loef, 1989). Teachers' mathematical knowledge for teaching was in part assessed by their ability to make distinctions between different types of addition and subtraction word problems, and their general knowledge of student strategies for solving addition and subtraction problems. The CGI researchers found that the students of teachers who exhibited more robust knowledge of different types of addition and subtraction problems and the types of student solution strategies scored higher in word-problem assessment. However, the difference between students' computational assessment scores revealed no difference between teachers with and without robust knowledge of different types of (a) addition and subtraction problems, and (b) student solution strategies.

Even (1993) investigated secondary teachers knowledge within in the content domain of functions. She used the constructs of PCK and the work of CGI to frame her study. In fact, Even's work incorporated many components that would later be discussed by Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Hill, Rowan & Ball, 2005) through the construct of MKT. Even used both questionnaires and follow-up interviews to gather data on secondary teachers' knowledge. Tasks included asking teachers to first provide a definition of function and then to provide an alternative definition for a student who struggled to understand the first definition. Also, Even provided examples of student work and asked teachers to determine if the student was right or wrong and why. These kinds of tasks allowed Even to evaluate teachers' subject matter knowledge and their knowledge of student solution strategies. Asking teachers for their definition of function, for example, allowed Even to determine the teachers' knowledge of function. In addition, asking for an alternative version of the definition provided Even the opportunity to examine teachers' content knowledge, by assessing the depth of their understanding of function, as well as teachers' knowledge of content and students, by assessing the repertoire of explanations they have for students.

2.3.3 Procedural and Conceptual Constructs of Knowledge

Other researchers (Eisenhart, Borko, Underhill, Brown, Jones, & Agard, 1993) studied teacher knowledge through the constructs of procedural and conceptual knowledge. Eisenhart et al. used Hiebert's (1986) definitions for conceptual and

procedural knowledge because these definitions are regularly used by the larger

mathematics education community as a means for communicating about these

constructs. Hiebert defines conceptual knowledge as "knowledge that is rich in

relationships" (p. 3), and procedural knowledge as knowledge of "the formal

language, or symbol representation system" and knowledge of "the algorithms, or

rules for completing mathematical tasks" (p. 6). Although the dichotomy between

these two types of knowledge is artificial, and the distinctions between procedural and

conceptual knowledge are certainly more interconnected in practice, it is useful for

discussion purposes to examine different types and qualities of knowledge.

Eisenhart et al. (1993) argued that the novice teachers in their study spent more

time teaching for procedural knowledge than conceptual knowledge. Eisenhart et al.

found that although the teachers in their study expressed interest in teaching for

conceptual knowledge, a number of factors made it more challenging for these

teachers to actually teach in a manner that would support the development of

conceptual knowledge. In particular, Eisenhart et al. suggested three factors which

limited teachers' abilities to teach for conceptual knowledge: (1) teachers' own limited

conceptual knowledge; (2) expectations placed on novice teachers by their cooperating

teachers and their school district influenced novice teachers' priorities toward

covering all of the procedural skills laid out in the curriculum and preparing students

for standardized tests before emphasizing conceptual knowledge; and, (3) novice

teachers were influenced to teach procedural skills first because this teaching

philosophy appeared to be supported in the department by cooperating teachers and by

administrators. As a result of the present study, I suggest another potential limiting factor – a teacher's prior experiences learning mathematics is likely to influence the type and quality of mathematics they work to develop in their students; Even (1993) also makes this argument.

2.3.4 Coordinating PCK, MKT, Procedural and Conceptual Knowledge

Fennema and Franke (1992) put forth the model shown in Figure 10 as a viable framework for illustrating the dynamic and integrated features of teacher knowledge. Specific features in Figure 10, such as *knowledge of mathematics* and *knowledge of learners' cognitions in mathematics*, represent the type of knowledge components originally mentioned by Shulman (1986) and later refined by Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Hill, Rowan & Ball, 2005) into the construct of mathematical knowledge for teaching. Inside teachers' context specific knowledge resides their understanding of the procedures and concepts for a particular mathematical domain and the ways in which they connect those procedure and concepts to other areas of the mathematics curricula – the 'big ideas'. Also included here are teachers' abilities to apply this knowledge to novel mathematical problems and teaching situations. Teachers' knowledge of student reasoning and development in relation to a particular mathematical topic is also interconnected with their context specific knowledge.

Figure 10: Model of Teacher Knowledge

Fennema & Franke's (1992) Model of Teacher Knowledge (p. 162)
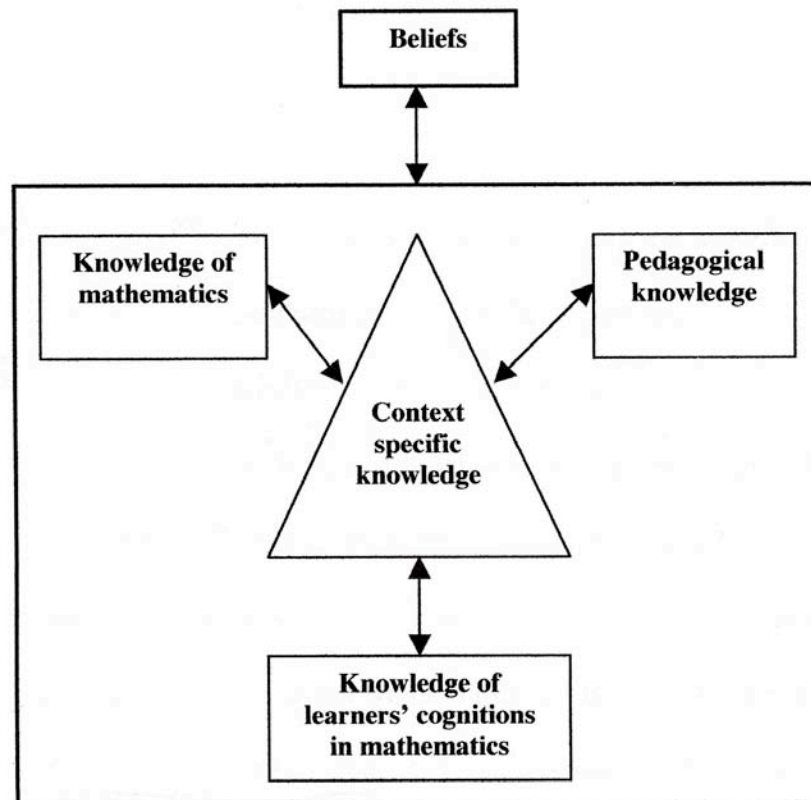
In addition, Fennema and Franke's framework presents a picture of how beliefs mediate the features of teacher knowledge. The issue of beliefs is addressed in this last subsection.

2.3.5 Teachers' Beliefs

The constructs of belief and knowledge are too closely connected to parse out in tidy categories, and often the words are used interchangeably. The research literature

(e.g., Thompson, 1992; Pajares, 1992; Kagan, 1992) suggests that knowledge may be a form of belief, but one that has been accepted by the larger community as shared truth. According to these researchers, in order to characterize something as knowledge, there are certain criteria, established by the research community, which must be met. Beliefs, however, are not necessarily accepted by the larger community and remain personal in nature. Thompson articulates two main distinctions between beliefs and knowledge. Beliefs "can be held with varying degrees of conviction" and "beliefs are not consensual" (Thompson, 1992, p.129).

In this study, I use knowledge to mean those beliefs that have undergone scrutiny by the established research community and are able to withstand that scrutiny. It is clear from the literature that beliefs and knowledge are intertwined, making it difficult to study one without necessarily looking at the other. However, the work of Ball and her colleagues (Ball, Lubienski, & Mewborn, 2001; Ball & Bass, 2003; Hill, Rowan, & Ball, 2005), CGI (Carpenter, Fennema, Peterson, & Carey, 1988; Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; Peterson, Fennema, Carpenter, & Loef, 1989; Fennema, Carpenter, Franke, Levi, Jacobs, & Empson, 1996), and Even (1993), provide a model for assessing teachers' mathematical knowledge for teaching against research literature on students' mathematical reasoning. For example, Ball and her colleagues, CGI, and Even used students' alternative ways of approaching mathematical tasks, as evidenced by the research literature, as a means for assessing teacher knowledge of content, and knowledge of content and students.

2.3.6 Summary

The work of Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Ball & McDiarmid, 1990; Hill, Rowan & Ball, 2005), CGI (Carpenter, Fennema, Peterson, & Carey, 1988; Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; Peterson, Fennema, Carpenter, & Loef, 1989; Fennema, Carpenter, Franke, Levi, Jacobs, & Empson, 1996), and Eisenhart et al. (1993) pertains to the mathematical knowledge necessary for teaching elementary school mathematics, and Even's (1993) study pertains to secondary teachers' concept of function. However, the models provided by these researchers are useful for thinking about the necessary knowledge for teaching introductory statistics courses. First, fusing the constructs of common content knowledge, specialized content knowledge, and knowledge of content and students/teaching with the constructs of statistical literacy, statistical thinking, and statistical reasoning, respectively, provide a framework for investigating TAs' statistical knowledge for teaching. Figure 11 provides a visual representation of this framework. I highly specify the framework referenced here in Chapter 6, because this framework in large part emerged from my analysis of the data.

Figure 11: Model of Statistical Knowledge for Teaching

Model for Statistical Knowledge for Teaching

Statistical Literacy Skills
Common Content Knowledge

a) Understand common statistical terms
b) Read and make sense of statistics in the media
c) Ability to be a critical consumer of statistics
d) Informal statistical inference skills

Statistical Thinking
Specialized Content Knowledge

a) Deep and well connected knowledge of introductory statistics material – descriptive & inferential
b) Consumer and producer of statistics – Design experiments, collect data, analyze data, draw

Knowledge of Content and Students

a) Alternative solution strategies
b) Common student reasoning paths
c) Common conceptual hurdles

Second, the work of Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Ball & McDiarmid, 1990; Hill, Rowan &

Ball, 2005), CGI (Carpenter, Fennema, Peterson, & Carey, 1988; Carpenter, Fennema,

Peterson, Chiang, & Loef, 1989; Peterson, Fennema, Carpenter, & Loef, 1989;

Fennema, Carpenter, Franke, Levi, Jacobs, & Empson, 1996), and Eisenhart et al.

(1993) provide a methodological model for assessing teacher knowledge of content

and students. Specifically, using hypothetical student work as the basis for task-based

interviews provides a means for assessing teachers' content knowledge and their

knowledge of content and students, as compared to the research community's

knowledge of content and students.

2.4 Chapter Conclusions

In conclusion, this chapter provides an overview of the theories and literature that

support my research. Radical constructivism provides a guiding philosophy, which

impacts my perspectives on how learning takes place, how teaching can best support

such learning, and the overarching purpose of statistics education. First, my

epistemological assumptions function on a meta-level, orienting my research design

and analysis in order to support the creation of a viable model for TAs' statistical

content knowledge. My review of the statistics education literature highlights the

importance of research on sampling concepts. In my review of the literature, I

synthesized the features necessary for building a profound understanding of sampling,

and identified types of student difficulties that emerge as students' statistical reasoning

develops. In addition, I identified a useful framework for investigating reasoning about

empirical sampling distributions (Shaughnessy et al. 2004a&b, 2005) and another for

investigating the relationship of statistical inference to sampling (Liu & Thompson,

2005). These themes, gleaned from the research literature, provided a structural

framework for engineering my research instruments and for orienting my data

analysis. This structural framework enabled me to create a viable model of TA

reasoning on a micro-level.

Second, a constructivist epistemology has implications to theories of teaching. If

an individual learns by constructing his or her own knowledge through experiences,

then a natural implication for teaching would be that instruction should begin with

students' understandings. Maher and Alston (1990) discuss the implications of

constructivism on teaching. They state,

> An important foundation for constructing even more complex systems of
> knowledge about teaching includes the building of systems of knowledge
> about the following:
>
> How children interpret the ideas in school mathematics;
> What kinds of strategies children invent and use; and
> How to interpret the kinds of errors children make.
>
> Attention to these behaviors better enables the teacher to aid the student in
> building more powerful constructions (p.150).

Although Maher and Alston's comments are in reference to K-12 grade students, they

are appropriate for college teaching as well. I believe that college students should also

be actively involved in constructing appropriate understandings of key probability and

statistics concepts. Further, I do not believe that most college students will leave

introductory statistics with meaningful and useful understandings of statistics if they

are only shown processes and procedures for finding solutions to routine problems.

Students need to be actively engaged in sampling activities, computer simulations, and

class discussions about the messiness of data collection and analysis of data in order to construct a deeper understanding and appreciation of the field of statistics. Pfannkuch (2005), and Cobb and Moore (1997) argue that students need to be able to think like statisticians, and I believe that active engagement in these types of activities can lead to such thinking. These are the core beliefs that guide my research.

The work of Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Ball & McDiarmid, 1990; Hill, Rowan & Ball, 2005), CGI (Carpenter, Fennema, Peterson, & Carey, 1988; Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; Peterson, Fennema, Carpenter, & Loef, 1989; Fennema, Carpenter, Franke, Levi, Jacobs, & Empson, 1996), and Even (1993) provide support for the position I outlined in the preceding paragraph. The work of these researchers melded with the research on student reasoning and the type of knowledge structures students need in order to develop statistical literacy and thinking skills provides a micro-level structure for researching TAs' statistical knowledge for teaching. Specifically, Ball and her colleagues, CGI, and Even provided two methodological considerations useful to this study. First, the focus on a specific mathematical content area allowed these researchers to develop a more robust characterization of teacher knowledge in the teaching of a particular topic. Second, using hypothetical student responses as a proxy for understanding TAs' content knowledge and knowledge of content and students proves to be a valid and reliable methodological tool. In addition, the research presented in this chapter on teacher knowledge provided an analysis tool for examining TAs' statistical knowledge for

teaching. In Chapter 3, I elaborate further about the ways in which I used the research

presented here to design this research study and analyze the data.

CHAPTER 3

RESEARCH DESIGN, METHODOLOGY AND ANALYSIS

As described in Chapter 1, the primary purpose of this study was to characterize

TAs' statistical knowledge for teaching sampling processes. In particular, I

investigated: 1) How TAs understand the concepts of sampling; 2) How TAs

conceptualize the relationship between sampling and statistical inference; 3) How TAs

relate sampling and statistical inference concepts to probability; and, 4) TAs'

knowledge of content and students. In this chapter I detail the research methodology,

design, and analysis for this study. This chapter is presented in five sections. In

Section 3.1, I provide a general overview of the research design, data collection

methods, and rationale for these methods. In Section 3.2, I discuss the research

instruments. In Section 3.3, I discuss participant selection. In Section 3.4, I discuss

how data analysis was conducted, including a detailed outline of the phases of the

analysis. In Section 3.5, I address issues of validity in this research project.

3.1 Research Methodology and Design

One of the primary purposes of this study was to develop a rich and detailed

understanding of TAs' subject matter knowledge of sampling. This goal, guided by a

constructivist epistemology implied building a viable model of TAs' reasoning about

sampling processes. For a constructivist this means composing a framework of TAs'

conceptions via my interpretations of TAs' spoken words and/or written work. The

second primary goal of this study was to investigate TAs' knowledge of content and

students, including common conceptual difficulties. In my investigation of TAs'

conceptions of sampling, the primary focus was in building a viable model of how

TAs perceive of sampling processes; it was not to suggest a system of knowledge that

TAs *should* have. However, in my investigation of the necessary statistical knowledge

for teaching and TAs' knowledge of content and students, it was my intention to build

a model for the types of knowledge that *need* to be well-developed in order to achieve

effective teaching. In order to achieve this goal, I grounded my investigation with

statistics education research centered on student reasoning. This approach was also

guided by a constructivist epistemology in the following sense: for a teacher to

effectively facilitate students' constructions of knowledge, the teacher must know

something about how his/her students learn.

3.1.1 Framing the Research Design: Interplay between this research study, a constructivist epistemology, and prior research in stochastics education and teacher knowledge

My review of the research on students' and teachers' statistical reasoning about

sampling is germane to my research goals and my guiding constructivist epistemology

because it enabled me to find (a) sampling tasks that have been shown to be valid and

reliable for constructing models of student and/or teacher reasoning, and (b)

frameworks for aiding in the preliminary analysis of TAs' reasoning. It is not

unreasonable to assume that TAs will have similar stages in their statistical

development as students of statistics and K-12 teachers, although it may be true that

TAs are already at a more advanced stage. However, this conjecture has not been

tested. Thus, tasks were reproductions of, or modeled after, many of the tasks

discussed in the literature review (see Chapter 2) in order to engage TAs in activities

and discourse that would support building viable models of TA knowledge.

In addition, the literature review of students' and teachers' statistical reasoning

pointed to common conceptual difficulties, and to knowledge structures that all adults

should have – statistical literacy, and knowledge structures *necessary* for statistics

teachers and statisticians – statistical thinking. These two features of my literature

review, coupled with the research on teacher knowledge, provided support for

developing a model of statistical knowledge necessary for effective teaching, and as a

means for assessing TAs' knowledge of content and students. Also, the research on

teacher knowledge provided a methodological approach of using hypothetical student

tasks as a means for investigating teachers' thinking about their students' learning.

Figure 12 represents my conceptual map for the interaction between a constructivist

epistemology, the research on student and teacher reasoning about sampling, research

on teacher knowledge, and how these elements frame my research design and analysis.

Figure 12: Methodological Model

Conceptual Map: how the background literature and guiding philosophy work together to inform my research design and analysis

Macro Level: Radical Constructivism
- Philosophy of learning – learner constructs his/her knowledge
- Impacts what it means to "know something" and how I investigate TAs' knowledge
- Impacts my view of the purpose of statistics education and what it means to engage in statistical activity
- Impacts my view on teaching practices and what constitutes statistical knowledge for teaching

Implication 1: Build a viable model of TAs' conceptions of sampling processes, not a model of what they *should* know
- Research design must provide rich sources of TAs' interpretations and reasoning

Implication 2: Teachers must know something about how students learn in order to facilitate construction of students' knowledge
- Research design must provide rich sources of TAs' interpretations of student thinking and use existing research on student reasoning as a basis for examining TAs' knowledge

Micro Level: Research on Stochastics Education and Teacher Knowledge
Implication 1:
- Research design – use task based interviews/surveys that have been shown in the past to be valid and reliable for measuring students' and K-12 teachers' conceptions of sampling
- Analysis – begin with Shaughnessy et al. (2004, 2005) framework for examining TAs' reasoning on experimental graphs and Liu and Thompson's (2005) framework for statistical inference tasks

Implication 2:
- Research design – use research methodology of CGI and Even (1993) by creating hypothetical student responses grounded in my synthesis of building blocks & common student difficulties
- Analysis – use the common building blocks, common misconceptions and conceptual frameworks as a means for comparison to TAs' interpretations of student reasoning.

3.1.2 Data Collection Methods

In order to achieve the goals outlined in the preceding paragraphs, I first used a task-based web survey followed by a series of three 80-minute interviews with a small subset taken from the larger group of survey participants. The surveys provided me the opportunity to gather, from a larger, more diverse group of TAs, general information about their educational backgrounds and certain limited information on how they reason in sampling contexts. Also, the surveys provided me the opportunity to find volunteer interview participants. The interviews supplied opportunities to follow up with TAs' interpretations and responses to the survey tasks and to continue to build a viable model of TAs' statistical knowledge for teaching by engaging TAs in (a) statistical conversations about sampling processes in a variety of contexts, and (b) conversations about the teaching and learning of sampling concepts. The series of three interviews with each interview participant is what allowed me to achieve a detailed, fine-grained analysis of these TAs' statistical knowledge for teaching sampling. Repeated interviews with the same TAs allowed for corroboration of findings. Taken together, the surveys and interview series enabled me to triangulate the data, because I was able to gather snapshots at different moments in time of the ways in which TAs reasoned about comparable sampling tasks and how they reasoned across a variety of sampling tasks.

In addition to the surveys providing baseline information about TAs' conceptions of sampling processes, TAs' responses informed the first interview questions.

Although tentative interview protocols were in place, in some sense interview tasks were modified in order to adjust to information on TA reasoning that emerged from the survey and needed to be further clarified or specified. For example, the second and third interviews were modified when the data from prior interviews suggested a compelling idea that needed further illumination in order to build viable models of TAs' statistical knowledge for teaching. The interview case studies that emerged in this study serve as exemplars for characterizing TAs' knowledge of sampling processes, and may provide statistics educators' insight into areas of undergraduate and graduate statistics education deserving careful attention.

3.2 Survey and Interview Instruments

Table 1 shows the sequence of data collection that took place during the fall of 2006. The tasks are located in the appendix. First I describe the survey tasks and follow with a description of the interview tasks.

Table 1: Data Collection Timeline

Events	Dates	Tasks
Survey	September 2006	Unusual Sample Task Prediction Task Real/Fake Task
Interview 1	October 2006	Discussion and follow-up questions to survey tasks
Interview 2	October 2006	Mystery Mixture Task Terminology
Interview 3	November 2006	Gallup Poll Task Background information

3.2.1 Survey Tasks

The Unusual Sample Task is borrowed from Watson (2004, p. 281), and was
discussed in Chapter 2. The Unusual Sample Task addresses the relationship between
sample size and sample variability and connections between probability and sampling.
The Prediction Task was borrowed from an interview task previously used on a
National Science Foundation (NSF) research grant, titled *Students' Conceptions of
Variability*; see Shaughnessy et al. (2004 a & b). This task assesses distributional
reasoning, knowledge of theoretical and empirical sampling distributions from a
known population, probability structures, and the balance between sample variability
and sample representativeness. The second part of the Prediction Task was modeled
after Even's (1993) work on teachers' pedagogical content knowledge of functions. As
I mentioned in Chapter 2, Even provided examples of student work and asked teachers
to respond to the student work. Thus, for the second portion of the Prediction Task, I
provided TAs with hypothetical student predictions, typical of student work as
evidenced by the Shaughnessy et al. study (2004a&b), and asked TAs to respond to
the student predictions. This portion of the task addresses TAs' content knowledge of
empirical sampling distributions, but in addition it provides an opportunity to learn
about TAs' knowledge of content and students. The Real/Fake Task is borrowed from
an interview task previously used in a NSF research grant, titled *Students'
Conceptions of Variability*; see Shaughnessy et al. (2004a&b). This task assesses TAs'
knowledge of empirical sampling distributions and variability within and between
sampling distributions.

3.2.2 Interview Tasks

The first interview was designed as a follow-up to the survey. First, I followed up on the Unusual Sample Task. Giving TAs an opportunity to discuss how they interpreted the task and share more detail about their thought processes provided more information from which to build a model of TAs' knowledge with respect to this type of task. It is important to determine how robust TAs' knowledge of sampling concepts are. In addition, by introducing a differing response used by other TAs, I attempted to introduce some cognitive conflict for TAs during the interview setting in order to get a sense for the depth of their understanding. Introducing a hypothetical TA or student response was also meant to assess TAs' knowledge of common conceptual hurdles in these problems. Given that students struggle with the role that sample size plays in sampling variability, it is important to establish whether or not TAs explicitly link this task to the idea of sampling bias and sample size and to recognize that this is a difficult concept for students. Therefore, I also asked TAs how they thought students might respond to the Unusual Sample Task and what types of difficulties they thought students might have.

Second, I followed up on the Prediction Task. TAs had the opportunity to provide their interpretation of the task and why they responded the way they did. In particular, research suggests that students often experience tension between representativeness and variability. By providing examples of student predictions and asking TAs to elaborate on their reasons for deciding whether a student prediction was reasonable or unreasonable, I was able to get a sense of TAs' knowledge of content and students.

Third, I followed up on the Real/Fake Task. This task addresses TAs understanding of empirical sampling distributions. I attempted to introduce cognitive conflict by adding opposing views of other TA responses. In addition, I asked an alternative version of the task, where TAs rated the four sampling distributions from most likely to occur during the experiment to least likely to occur. The alternate phrasing allowed me to determine if changing the wording of the problem made a difference in how TAs interpreted the problem and how they would make their identifications. Finally, I asked TAs to discuss the types of difficulties students might experience with the Real/Fake Task, as well as what solution strategies they might employ.

The second interview provided an additional opportunity to build a model for thinking about TAs' statistical knowledge for teaching sampling. The Mystery Mixture Task is borrowed from a NSF research grant, titled *Students' Conceptions of Variability*; see Shaughnessy et al. (2004 a & b). I asked TAs to estimate the number of red candies in a jar of red and yellow candies. This task again addresses TAs' understanding of empirical sampling distributions and their relation to confidence intervals and statistical inference. In addition, during the second interview I asked TAs to discuss statistics terminology and how they might explain certain definitions and concepts to students. This allowed me to gain a deeper sense for how TAs thought about particular statistical concepts and their knowledge of how to introduce these topics to students.

The Gallup Poll Task was introduced during the third interview and was borrowed from Liu and Thompson (2005). This task addresses TAs' knowledge of sampling and its relation to confidence intervals, including the concepts of confidence level and margin of error. I also provided hypothetical student responses in order to assess TAs' knowledge of different interpretations of confidence interval and their understanding of students' differing perceptions of margin of error and confidence level. Finally, during the third interview I asked TAs to discuss how they thought of good teaching. Specifically, I asked TAs to describe (1) their past learning experiences; (2) what they liked or did not like about their past teachers; (3) influential teachers; and, (4) what, for them, made a teacher influential.

3.2.3 Summary

To summarize, the research instruments were designed to engage TAs in activities and discourse that would support the development of a model of their subject matter knowledge of sampling processes and their knowledge of content and students. The survey and interview tasks were developed with two main purposes in mind. First, tasks were designed to illicit discussion around the core conceptual components necessary for an understanding of samples, sampling distributions, and the relationship between sampling and statistical inference. Second, tasks were designed to assess TAs' knowledge of, and to promote discourse around, the common conceptual hurdles that students experience as they learn about ideas of sampling. Table 2 provides a map

for how each survey and interview task addresses the particular conceptual building

blocks and common student difficulties that were identified in Chapter 2.

Table 2: Correspondence between tasks and knowledge components

Conceptual Building Blocks	Corresponding Conceptual Difficulties	Corresponding Survey & Interview Tasks
Definition of Sample	Difficulty with colloquial versus statistical use of the term	The Unusual Sample Task & Terminology
Proper Sampling Methods	Difficulty with random selection and/or recognizing bias in the sampling process	The Unusual Sample Task & Terminology
Coordinate multiple attributes of distributions	Difficulty coordinating multiple attributes – overly focused on center, shape or spread	The Prediction Task, Real/Fake Task, & Mystery Mixture Task
Sampling Distributions	Difficulty with empirical versus theoretical models – bounded variability	The Prediction Task, Real/Fake Task, Mystery Mixture Task, Terminology, Unusual Sample Task, & Gallup Poll Task
Relationship between sampling variability and sample size	Difficulty recognizing as sample size increases sample variability decreases – no image of repeating the sampling process	The Unusual Sample Task
Role of sampling in statistical inference	Difficulty interpreting confidence intervals with long-term relative frequency perspective and an image of a distribution of sample statistics	The Mystery Mixture Task, & Gallup Poll Task

3.3 Survey and Interview Participants

3.3.1 Survey Participants

Sixty-eight graduate teaching assistants from 18 universities around the United

States participated in the task-based web survey. The survey participants comprise a

convenience sample in which participation was voluntary. Participant eligibility

required TAs be enrolled in a graduate statistics, mathematics, mathematics education

or related graduate program. Universities were targeted for solicitation wherever contacts existed between the researcher and the university.

In the Fall of 2006, TAs were sent a link to the web survey via email. Approximately one week after the initial solicitation email, TAs received one reminder notice. TAs who participated in the survey were entered into a raffle to win an iPod®. Survey participants were required to leave their name and email in order to be entered into the raffle, but this information was kept confidential, and identifying information was removed from survey responses. Table 3 and Figure 13 provide demographic summary information on the survey participants. There were approximately equal numbers of male and female participants. Notice that a 'typical' survey participant was in his/her 20's, and spoke English as a first language.

Table 3: Demographic information on survey participants

English as a Second Language		Gender	
Yes	14 (21%)	Male	36 (53%)
No	54 (79%)	Female	32 (47%)
Total	68 (100 %)	Total	68 (100%)

Figure 13: Age Distribution for Survey Participants
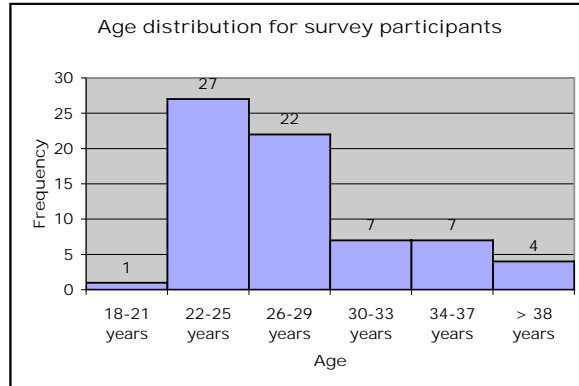
Table 4 provides information on the survey participants' mathematics and statistics background. Approximately 80% of survey participants were enrolled in a graduate statistics program at the time of the survey and 97% had taken at least one graduate statistics course. In addition, 85.3% of survey participants had taught their own section or worked as a recitation instructor for an introductory college statistics course.

Table 4: Mathematics and statistics background of survey participants

Undergraduate Degree		Current Field of Study		Number of Undergraduate Statistics Courses		Number of Graduate Statistics Courses	
Statistics	12 (17.6%)	Statistics	54 (79.4%)	0 courses	6 (8.8%)	0 courses	2 (2.9%)
Mathematics	39 (57.4%)	Mathematics	7 (10.3%)	1-3 courses	30 (44.1%)	1-3 courses	19 (27.9%)
Mathematics Education	1 (1.5%)	Mathematics Education	3 (4.4%)	4-7 courses	18 (26.5%)	4-7 courses	9 (13.2%)
Other	16 (23.5%)	Other	4 (5.9%)	8-11 courses	10 (14.7%)	8-11 courses	15 (22.1%)
				More than 11 courses	4 (5.9%)	More than 11 courses	23 (33.8%)

The background information on these survey participants suggests that I reached my targeted population, in the sense that the majority of my participants were graduate students with at least one graduate statistics course, and who had some experience teaching undergraduate introductory statistics.

3.3.2 Interview Participants

The interview participants consisted of a subset of five TAs from the larger survey population. All interview participants were from the same university. As with the survey participants, the interview participants comprised a convenience sample. Interview participants volunteered to participate and received a $15 stipend per interview for their time. The eligibility requirements for interview participants were (a) having taken at least one graduate statistics course, and (b) having taught introductory college statistics at least one time. Interviews took place over a two-month period during the Fall of 2006. Video and audio recordings were conducted with two interview participants. Three interview participants consented only to audio recordings. Interview participants were given pseudonyms, and the video and audio recordings were kept locked in the researchers filing cabinet.

In order to preserve the confidentiality of the interview participants, I use pseudonyms: Amanda, Sandy, Joe, Andy, and Sam[14]. Further, I do not match up detailed information about the majors and backgrounds for each TA. Rather, I provide

---

[14] In the analysis chapters, Chapters 4, 5, and 6, Sam enters the discussion only periodically. Sam is often omitted from the discussion because he did not articulate his thoughts as clearly as the other TAs and hence did not add much depth to the discussion. It is worth pointing out that Sam spoke English as a second language and tended to be shy about speaking; therefore, there was not enough substance from his interviews to warrant including him in every discussion.

background information as a group for the five TAs[15]. The demographic break down

for the interview participants is as follows: two females and three males, two

international students who spoke English as a second language, and an age range from

27 to 38 years of age. Table 5 shows the mathematics and statistics background and

fields of study for the five interview participants.

Table 5: Background Information of Interview Participants

Program		Major		Number of undergraduate statistics courses		Number of graduate statistics courses	
Ph.D.	3	Statistics	3	0 courses	2	1 course	1
Master	2	Mathematics	1	1-3 courses	1	8-11 courses	1
		Mathematics Education	1	4-7 courses	2	More than 11 courses	3

The teaching background for the five interview participants is as follows: two of the

TAs had taught the first quarter of introductory college statistics[16] only one time, one

TA taught the first quarter of introductory statistics three times, and two TAs had

taught both quarters of the introductory college statistics course multiple times. In

addition, one TA taught a 400/500 level statistics for engineers course one time.

3.4 Data Analysis

In this section I describe my data analysis methods. The analytic approach I

employed is consistent with grounded theory methods (Glaser and Strauss, 1967), in

---

[15] Interested researchers can contact the author of this dissertation directly should that information be necessary for some reason.

[16] The university where this research took place was on the quarter system. The first term of introductory statistics at this particular university covered introductory probability, Binominal and Poisson distributions, descriptive statistics, the Central Limit Theorem, and basic confidence intervals. The second term continues work with statistical inference – confidence intervals and hypothesis testing.

which data analysis is an iterative process where hypothesis are generated, reflected upon, and modified over several cycles until increasingly stable and viable hypotheses emerge. Using this approach to data analysis I: (1) reviewed the data and formulated initial descriptions and hypotheses of TAs' understandings in each task; (2) tested the viability of my initial hypotheses by searching the data for conflicting or supporting evidence; and, (3) revised my initial hypotheses on the basis of the conflicting or supporting evidence gathered in my searches. By reiterating this cycle of investigation, my hypotheses developed into stable and viable models for characterizing TAs' statistical knowledge for teaching.

The data corpus for analysis included 68 survey responses and 15 (1-1.5 hour long) video and/or audio taped interviews, three interviews for each of the five TAs. First, I evaluated TAs' performance on the survey by scoring their responses to tasks against a pre-established set of criteria grounded in a normative response[17]. Second, I examined the survey data and formulated initial impressions concerning TAs' conceptions of sampling ideas and knowledge for teaching statistics. Following my initial examinations of the surveys, I triangulated with the interview data sources in order to test initial impressions gathered in the survey data. In my initial examination of the interview data, I looked for evidence that would refute or support my initial impressions of the survey data. I also created new impressions and conjectures of TA reasoning from initial reviews of the interview data because it was more detailed and

---

[17] The normative criteria for assessing the survey data are discussed in the conceptual analyses of the tasks in Chapters 4 and 5.

richer than the survey data, and provided a different insight into TA thinking. As I iterated this procedure of reviewing the survey and interview data with a critical eye toward my conjectures, I was able to locate themes in the data and to build stable hypotheses.

3.4.1 Level 1: Preliminary Examination of Survey and Interview Data

In my first pass of the survey data, I scored surveys based on criteria established prior to the distribution of the surveys. The criteria were established by comparison of normative response to non-normative responses. In my second pass through the survey data, I categorized TAs' responses on the Prediction and Real/Fake Tasks according to the conceptual framework of Shaughnessy et al. (2004a&b, 2005), making note of cases that did not fit into this framework. Cases that did not fit into this framework were reviewed again in order to create new categories of responses and to begin the initial revision of the framework of Shaughnessy et al. for modeling TAs' statistical reasoning. In my third pass through the survey data, I looked for themes in TAs' reasoning and compared TAs scored responses to types of reasoning employed.

My first pass of the interview data consisted of writing a reflection of what transpired immediately following each interview. Later, I reviewed the video/audio data, taking notes of interesting excerpts and again writing a summary of what transpired during the interviews. I then compared my initial reflections with my first review of the interview data. During this comparison I looked for places of agreement

or disagreement in my initial reflections after the interview and my observations during the first viewing of the interview data.

3.4.2 Level 2: Transcription Analysis

At this level of analysis I created verbatim transcripts of all the interview data. As I created transcripts, I highlighted excerpts that appeared to express direct evidence of (a) TAs' reasoning about a sampling task, and/or (b) how they thought about student learning. I highlighted places where there appeared to be miscommunication so that, if possible, I could follow up on those conversations in subsequent interviews. I took extensive notes for where there appeared to be agreement or disagreement in my application and refinement of the framework of Shaughnessy et al. (2004a&b, 2005) and/or Liu and Thompson's framework[18], and where appropriate I compared these notes with the categories of reasoning I created from the survey responses.

3.4.3 Level 3: Detailed Coding Analysis

Upon completion of the transcripts and first series of extensive notes about TAs' thinking, I read through the transcripts again. During this time I looked for evidence that would confirm or refute the categories I had created for TAs' reasoning and for TAs' knowledge of content and students. I refined my categories for characterizing TAs' statistical knowledge for teaching. I compared reasoning across different tasks and I compared the reasoning of different interview participants. I created concept

---

[18] Shaughnessy's framework was applied to the Prediction, Real/Fake and Mystery Mixture Tasks. Liu and Thompson's framework was applied to the Gallup Poll Task.

maps for charting their reasoning paths and sought to explain reasons for apparent contradictions in TAs' thinking. Following this, I wrote preliminary summaries of the types of thinking that appeared to be emerging and how that thinking might best be modeled. In these summaries I included both broad and detailed categories of reasoning, similarities and/or difference in survey and interview codes, and a comparison of types of reasoning among the interview participants.

3.4.4 Level 4: Chronicling Emerging Themes

At this point, I was ready to tell the story of how the survey and interview participants reasoned about the sampling tasks, and how they thought about student learning. I reviewed my prior layers of analysis, fine-tuning my coding scheme as I progressed through this review and reflection period. I sought to clarify and refine the summaries I created in Level 3. At this point, compelling themes emerged about how TAs: (1) reasoned with experimental sampling distributions; (2) reasoned about sampling and statistical inference; and, (3) reasoned about student learning and teaching in these contexts. In particular, the following themes form the basis of the next three chapters of analysis: (1) TAs experienced tension reasoning with experimental data; (2) TAs reasoned about sampling experiments and statistical inference on a spectrum ranging from no conception of repeated sampling to strong conceptions of repeated sampling; and, (3) TAs displayed limited statistical knowledge for teaching as a result of limitations in their subject matter knowledge and/or their knowledge of content and students.

3.5 Issues of Validity

In order to ensure that my analysis yielded credible results, issues of validity have remained in the forefront for both the design and analysis of this research project. First, collecting multiple forms of data – survey and interview, and conducting multiple interviews with the same participants over an extended period of time, enabled the collection of rich and detailed information on TAs that should serve to triangulate data and corroborate findings. Second, with each level of analysis I looked for alternative explanations for TAs' responses, and for contrasting cases that did not fit within the coding scheme. As rival explanations entered the scene during the analysis, I revised both the coding scheme and subsequent interviews in order to test alternative explanations.

OVERVIEW OF CHAPTERS 4, 5, & 6

Chapters 4, 5, and 6 comprise my analysis of the data. These three chapters address the two overarching goals of this study – an investigation of TAs' statistical content knowledge and their knowledge for teaching statistics. These three Chapters specifically address the model of statistical knowledge for teaching that I briefly outlined in Chapter 2 (recall Figure 11). Chapters 4 and 5 address my investigation of TAs' statistical content knowledge. I detail my analysis of TAs' thinking and reasoning on the survey and interview tasks. In particular, I discuss two themes that emerged from my analysis of the data: 1) TAs in this study appeared to experience tension, which they could not always resolve, between their knowledge of theoretical models and experimental data; and, 2) TAs in this study appeared to have different ways of interpreting sampling and/or confidence interval problems. In Chapter 4, I discuss the first of these themes, characterizing how these TAs thought about experimental situations and the ways in which they used information from experimental data to answer statistical questions. In particular, I discuss tensions that TAs appeared to experience in their expectations for experimental sampling distributions. In Chapter 5, I discuss the second theme, providing my interpretation of how these TAs thought about ideas in sampling and statistical inference. In particular, I discuss the ways in which probability entered into TAs' reasoning about sampling and statistical inference tasks.

In Chapter 6, I discuss the implications of TAs' subject matter knowledge and knowledge of content and students on their statistical knowledge for teaching. This chapter is presented in three sections. In Section 6.1, I provide a framework for necessary statistical knowledge for teaching sampling. In Section 6.2, I compare my model of TAs' interpretations of sampling ideas (as illuminated in Chapters 4 and 5) with those advocated by the statistical community. I argue that if TAs' thinking integrates multiple attributes of a distribution and if their reasoning about sampling ideas agrees with the norms set by the statistics community, then they are more likely to have a robust knowledge for teaching statistical ideas. In this section, I also discuss another component of TAs' statistical knowledge for teaching – TAs' knowledge of content and students[19]. After discussing TAs' statistical knowledge for teaching, I shift to the final section, Section 6.3, TAs' beliefs about teaching and learning. Although the focus of this dissertation study is about TAs' statistical knowledge for teaching, not TAs' beliefs about teaching, there is too much overlap between these two constructs not to address TAs' beliefs. During the interviews, TAs overwhelmingly discussed their beliefs about how statistics should be taught, or about their view on how students learn. Thus, in this section I present a brief argument for the ways in which the type and quality of TAs' subject-matter knowledge, along with their own experiences learning mathematics and statistics, appeared to influence their beliefs about how statistics should be taught and how they thought about student learning.

---

[19] Knowledge of content and students is one component, used in the manner of Ball (2005).

CHAPTER 4

TENSIONS TAs' EXPERIENCED BETWEEN THEORETICAL MODELS AND EXPERIMENTAL DATA

The purpose of this chapter is to highlight the first of two significant themes that emerged in my analysis of the data on TAs' statistical knowledge. This theme relates to how these TAs reasoned about experimental data. In this chapter I investigate how TAs understood probability distributions, sampling distributions and how they applied their knowledge of theory to experimental data. In particular, I observed that TAs appeared to experience tensions between theoretical models and experimental data, which they could not always resolve. The Prediction Task, Real/Fake Task, and Mystery Mixture Task (see appendix for tasks) required TAs to make decisions or predictions from, or about, experimental data. These tasks were used in this study to investigate TAs' knowledge in five key component areas: (1) measures of center, (2) measures of spread/variability, (3) measures of shape, (4) balance between sample variability and sample representativeness, and (5) the concept of sampling distribution and distinction between theoretical and empirical sampling distributions. Recall that I identified these five key component areas during my review of the literature as necessary building blocks for a robust and thorough understanding of sampling processes (see Figure 2, Chapter 2).

This chapter is presented in two sections – one for the Prediction and Real/Fake Task, and one for the Mystery Mixture Task. In Section 4.1, I provide a conceptual

analysis and framework for the Prediction and Real/Fake Tasks, followed by an interpretation of TAs' reasoning via the framework. In Section 4.2, I provide a conceptual analysis and framework for the Mystery Mixture Task, followed by an interpretation of TAs' reasoning. The analysis of TAs' reasoning in these sections highlight the key theme addressed in this chapter. It is important to note that the conceptual analysis and framework for each task, partially built on existing research literature, developed and emerged from my analysis of the data. The conceptual analyses and frameworks are the means through which I illustrate how these TAs reasoned about sampling processes, and thus constitute an end product to this study.

4.1 Prediction & Real/Fake Tasks

I begin with a conceptual analysis of the Prediction and Real/Fake Tasks and follow with a discussion on TAs' reasoning about these tasks. The discussion on TAs' reasoning highlights the central theme of this chapter – the difficulties TAs experienced in reasoning with experimental data.

4.1.1 Conceptual Analysis & Framework for the Prediction Task

The Prediction Task first appeared on my survey instrument and then formed the basis for follow-up questions in the first interview with TAs. The task is shown in Figure 14.

Figure 14: Prediction Task

PREDICTION TASK

A jar contains 1000 candies, 750 are red and 250 are yellow. The candies are mixed well. Suppose that you pull a random sample of 10 candies from the jar, record the number of reds, put the candies back in the jar and mix them up. Suppose you do this 50 times. How many times out of 50 do you think you would get a handful of 10 candies with:

Number of Red Candies in Handfuls of 10	Prediction
0 red candies	
1 red candies	
2 red candies	
3 red candies	
4 red candies	
5 red candies	
6 red candies	
7 red candies	
8 red candies	
9 red candies	
10 red candies	
Total	50

This task is a sampling task in which the underlying population proportion is known and the interest is in making a prediction about what is likely to occur in 50 samples of size 10. Assuming that a TA is highly likely to recognize the ratio of red to yellow candies, there are then three probable paths for making a prediction: (1) using only proportional reasoning and arguing based solely on the attribute of center; (2) using an informal distributional argument with at least two of the three attributes of the distribution – center, spread and shape; and, (3) connecting the sampling problem to the underlying probability structure and using a formal probability distribution

argument. To reason using a formal or informal probability distribution argument requires an understanding of the underlying proportion in order to appropriately anchor the distribution at its center. Figure 15 shows my conceptual framework for reasoning about the Prediction Task.

Figure 15: Conceptual Framework for the Prediction Task
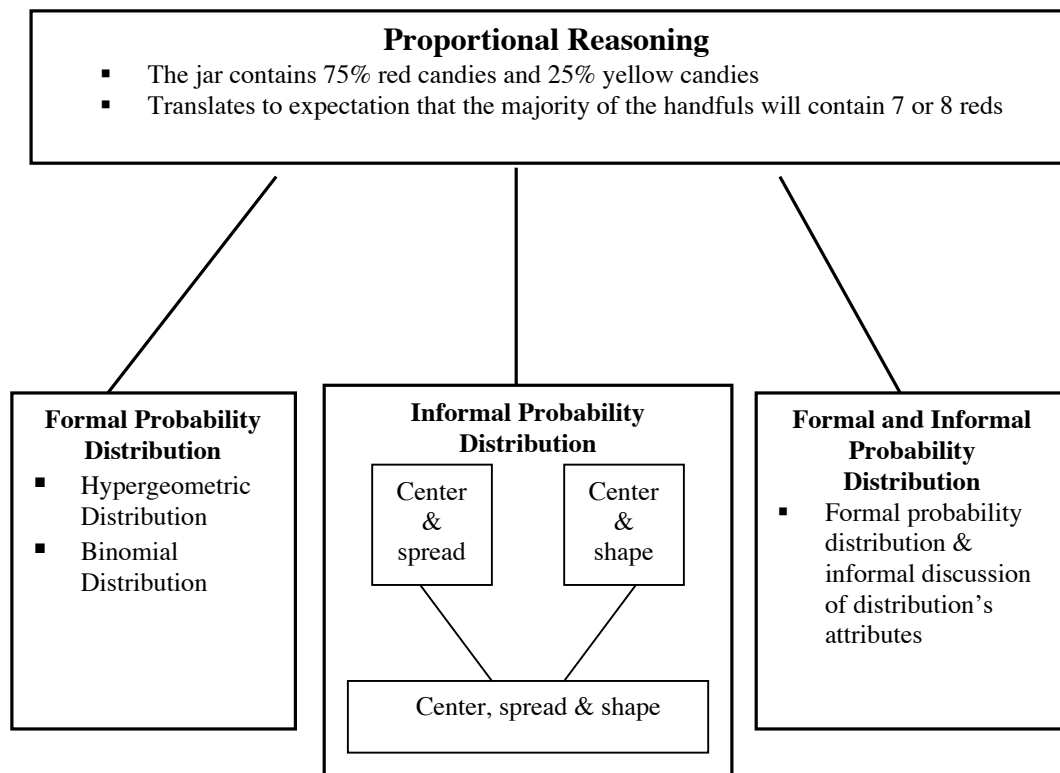
In the following sections I elaborate on the types of reasoning outlined in this conceptual framework for the Prediction Task. For each of these types of reasoning I describe the reasoning paths and arguments that TAs employed as a result of a thought experiment about the task. Following a deeper articulation of Proportional, Informal Probability Distribution, and Formal Probability Distribution reasoning, I discuss a

100

continuum scale on which these types of reasoning can be applied to resolve the tension between a theoretical model and experimental data.

Proportional Reasoning Argument

Attending mostly to the ratio of red to yellow candies, a proportional reasoning argument implies that the jar is 75% red. The ratio of red to yellow candies provides an indication that most of the samples would be expected to contain between seven and eight red candies – the center for this distributions. The frequency with which one predicts values at seven and eight red candies is a function of how much variability one expects in the 50 trials. So although with a proportional reasoning argument explicit attention is on the ratio of red to yellow candies, there is some implicit acknowledgement of variation. The amount to which variability from trial to trial is acknowledged could range from no acknowledgement (prediction for 50 trials is stacked at seven and eight red candies only) to acknowledging every possible outcome (placing at least one occurrence in each outcome, 0 to 10 red candies).

Informal Probability Distribution Argument

Reasoning with explicit attention to the ratio of red to yellow candies and at least one additional attribute of the empirical sampling distribution, spread, and/or shape, forms the basis for an informal probability distribution argument as a result of the thought experiment. Here variability becomes explicitly acknowledged and takes the form of a discussion about shape and/or spread of the empirical sampling distribution. A shape argument might include a discussion that the distribution has a particular

shape – e.g., left-skewed or symmetric around the center. A shape argument may also include a discussion about sample to sample variability, or the variability in the frequencies for the experimental sampling distribution (e.g., the change in vertical height of the bars of the experimental sampling distribution histogram from four red candies to five red candies for instance). A spread argument might include a discussion about the range of likely outcomes, standard deviation, or a focus on the left and/or right ends of the distribution.

Formal Probability Distribution Argument

If during the thought experiment the TA connects the sampling task to a theoretical probability distribution, then the hypergeometric or binomial are likely candidates to serve as a model distribution in the candy jar context. In a hypergeometric distribution, three assumptions must be satisfied: (a) the population is finite; (b) each element in the population can be characterized as a success or a failure; and, (c) a sample of size *n* is selected without replacement in such a way that each element is equally likely to be chosen. In the Prediction Task, the population of 1000 is clearly finite, and a success can be characterized as drawing a red candy and a failure as not drawing a red candy. When a sample of size 10 is picked, it can be assumed that each candy was equally likely to be chosen (since the jar was well mixed and candies were randomly selected), and the 10 candies are picked without replacement (so we are not picking the first candy out and replacing it before picking the second candy). With the assumptions of the hypergeometric model satisfied, the formula shown in Equation 1 can be applied,

where N is the size of the population, n is the sample size, M is the number of successes in the population, and x is the number of successes in the sample.

Equation 1

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Equation 1 expresses the theoretical likelihood of obtaining a sample, containing x successes from a population containing M successes. Finally, in order to find the expected value for zero red candies through ten red candies for 50 samples of size 10, one multiplies the probability for a single sample by 50. Table 6 shows the approximate distribution of reds according to the hypergeometric probability distribution. The values in the table are rounded to the nearest whole number because in the context of the problem a handful either contains x red candies or it does not (a handful of x red candies cannot happen a half time, for example).

Table 6: Hypergeometric Probability Distribution

Number of reds in a handful of 10	Expected Number of Occurrences
0	0
1	0
2	0
3	0
4	1
5	3
6	7
7	13
8	14
9	9
10	3

If instead a connection is made to a binomial probability distribution, then four assumptions must be met: (a) the population might be infinite, but the experiment consists of a specified number of individual trials; (b) each trial results in a success or a failure; (c) each trial is independent so that an outcome on one trial does not influence the outcome of another trial; and, (d) the probability of success on each trial is constant. In the Prediction Task, the population is finite, but the size of 1000 is large and the experiment consists of exactly 50 trials. Further, one could imagine the experiment such that each trial consists of a sample of size 10, where each of the 10 candies is picked out one at a time, its color is noted, and then the candy is replaced before picking out the next candy. With this image, each trial is independent, and the process of picking out one candy at a time creates a set of 10 smaller trials in each one of the larger trials, where the probability of success from trial to trial is constant. Alternatively, one could imagine that picking out all 10 candies, one at a time without replacing them, will not dramatically alter the probabilities given that the handful of 10 is so small compared to the population of 1000 candies. With the assumptions of the binomial satisfied (or approximately satisfied), the formula shown in Equation 2 can be applied, where n is the number of trials, x is the number of successes (red candies) in n trials, and p is the probability of success.

Equation 2

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Equation 2 expresses the theoretical likelihood of obtaining x successes from a

population with probability of success p. Using the binomial probability distribution

model the predicted values match the values of the hypergeometric probability

distribution model to the first decimal place. Thus the whole number values provided

in Table 1 from the hypergeometric model agree with the whole number values

provided by the binomial model. Regardless of choosing the hypergeometric or the

binomial probability distribution models, the probability function yields a distribution

with approximately 54% of occurrences at the center of the population (seven and

eight red candies), a shape that is left-skewed, and an interval range of seven units

(i.e., an expectation that handfuls will contain between four and ten red candies).

These three attributes provide a theoretical basis for an expected center, shape, and

spread for the data set.

Connecting Theoretical Expectations to Experimental Data

Regardless of the reasoning strategy, a TA must consider sample to sample

variability and/or variability between empirical sampling distributions in order to

understand what is likely to happen during the course of the experiment. That is,

knowledge of the concept of bounded variability (in the sense of Saldanha &

Thompson, 2003) – that balance between sample representativeness and variability (in

the sense of Rubin et al., 1991) – is necessary for reasoning about experimental data.

This notion may not be part of a TAs' conceptualization, and even if this

conceptualization is present it may be misconstrued. That is to say, a TA may not

expect any variability among outcomes, or have a sense for variability that is not bounded, and may not be particularly surprised by unusual outcomes. Thus, there is a continuum with which one can think about the variation in outcomes from one trial to the next. For example, in a proportional reasoning argument attention is focused on the ratio of red to yellow candies, but how does that translate into how many outcomes are placed at seven and eight red candies? Similarly, the expected outcomes provided in Table 1 are based on a theoretical model and in an experimental situation there will be some variability from one trial of size 10 to another trial of size 10. The tension resides in not knowing how much variability to expect from one trial to the next. If a TA does not have a robust knowledge of the concept of bounded variability, this tension will not be resolved. Questions that may follow when making predictions about experimental data from a theoretical model or a known population proportion include:

- What would an unusual prediction look like?
- How much variability can be expected in the percentage of occurrences at the center?
- How much variability can be expected in the left-skew shape?
- Will the shape of the graph be 'smooth', or will there be variability in frequency from one outcome to the next?
- How much variability can be expected in the range?
- How narrow or wide could the range be before it is considered unusual?

Answers to these questions are likely to be person specific. A TA's understanding of

an unusual trial of 50 is likely to reside on a continuous scale ranging from an exact to

inexact mapping with the theoretical probability distribution model. Figure 16

provides a representation of such a scale. An inexact mapping would indicate that the

TA's expectations of the experimental data differ considerably from the theoretical

model in shape, center and/or spread. An exact mapping would indicate that a TA's

expectations of the experimental data match exactly with the theoretical model in

shape, center, and/or spread. If a TA tends to believe that the experimental data will be

similar to the theoretical model, but expects some variability, then such an image

would fall closer to the right on the scale represented in Figure 16.

Figure 16: Continuum scale for describing variability in experimental data

Scale for Thinking about Variability in Experimental Data

Inexact Mapping		**Exact Mapping**
Expectation that anything can happen in the experiment – Unbounded variability • Expects wide or narrow interval range (spread changes) • Expects variable frequencies (shape changes) • Expects location of center to change	Along the continuum, expectations for the center, shape, and/or spread change (left to right) from more severe to less severe deviations from the theoretical model.	Expects little to no variation. Predictions match exactly to the theoretical model (center, shape, and spread)

Further, a TA's interpretation is likely to be based on a number of factors including their knowledge and understanding of the theoretical distribution, their intuition, and their experiences working with experimental data.

In order to set up criteria for which to compare and contrast TA's predictions, I returned to the underlying hypergeometric probability distribution for this experiment. A re-examination of the distribution in Table 1 suggests that: (1) approximately 27/50 or 54% of the handfuls contain seven or eight red candies; (2) approximately 86% of the handfuls contain between six and nine red candies; and, (3) the outcomes range from four to ten red candies. Yet, in experimental situations there will be some variability in the distribution; thus, I expect experimental data to look similar to, but not exactly like the theoretical model. Experimental data tend to have some variability in frequency rather than nice smooth shapes, and in this context we expect whole number values rather than fractional values as outcomes. Thus, the underlying probability structure and the context provided a natural set of criteria from which to compare and investigate TAs' predictions. Using the predicted center, shape, and spread provided by this distribution, I created a set of four criteria (built partially from criteria established by Shaughnessy et al., 2004 a & b) for evaluating how exact to inexact a TA's prediction was. Since the theoretical model provides an expected 27 outcomes at seven and eight red candies, fewer than 20 or more than 34 outcomes at seven and eight red candies produce graphs too uniform (less than 20) or too narrowly focused on center (more than 34) and thus, toward the inexact end of the continuum. These types of graphs would be more unusual. Between 24 and 30 outcomes at seven

and eight red candies is toward the exact end of the continuum and 20 to 34 outcomes

is toward the middle of the continuum. See Figure 17.

Figure 17: Criteria for Assessing Prediction Task

Prediction Task: Four Criteria for Assessing TAs' Predictions
- Whole number versus decimal predictions.
- Percentage of outcomes centered at 7 and 8 red candies.
  - Less than 20 (40%) or more than (34) 68% of the outcomes placed at seven or eight red candies is an inexact interpretation for experimental data.
  - Between 20 (40%) and 34 (68%) of the outcomes at seven or eight red candies would be more towards the middle of the continuum.
  - Between 24 (48%) and 30 (60%) of the outcomes at seven or eight red candies would be more towards the right of the continuum, exact end of the continuum.
- Interval range for the 50 trials of ten (spread)
  - An interval length of three or less, or nine or more would fall at the far left (inexact) end of the continuum. It suggests an 'anything goes' perspective (nine or more), or a narrow focus on center perspective (three or less). No sense of bounded variability.
  - An interval length of four, five, or eight would fall more in the middle of the continuum.
  - An interval length of six or seven would be at the exact end of scale.
- Shape of the distribution for 50 trials of size ten
  - The greater the variability in frequencies from one outcome (x-value) to the next, the farther left the prediction is on the continuum. In particular, a drop in frequency of more than 9 units from 7 to 6 red candies or from 8 to 9 red candies would be considered a significant deviation from the shape of the theoretical. Also, predicting similar outputs for 6, 7, 8, and 9 red candies would produce a shape too uniform. For example, if a TA predicted 10 handfuls with 6 red and 9 red and 12 handfuls with 7 red and 8 red. That is, approximately 80-92% of handfuls should fall within 6 to 9 red candies, less or more than this will create a more unlikely shape/spread.

I now turn to a conceptual analysis of the Real/Fake Task. In part, the analysis for the Prediction Task still holds for an examination of the Real/Fake graphs because the experimental situation is the same. The underlying probability structure is the same and the task could be examined using such a structure, or using less formal arguments based on one or more attributes of the distribution. Yet, the Real/Fake Task forces an examination of experimental data, so the continuum discussed previously is especially relevant in this situation.

4.1.2 Conceptual Analysis & Framework for the Real/Fake Task

The Real/Fake Task is an extension of the Prediction Task. The experimental situation is the same – there is a population of 1000 candies, 750 red and 250 yellow. The experiment consists of 50 trials of 10, but in this task TAs were presented with the graphs of four experimental sampling distributions and asked to identify which graphs were real (produced by simulation or performing the experiment) and which graphs were fake (produced by a student who did not do the experiment) (see Figure 18).

Figure 18: Real/Fake Graphs
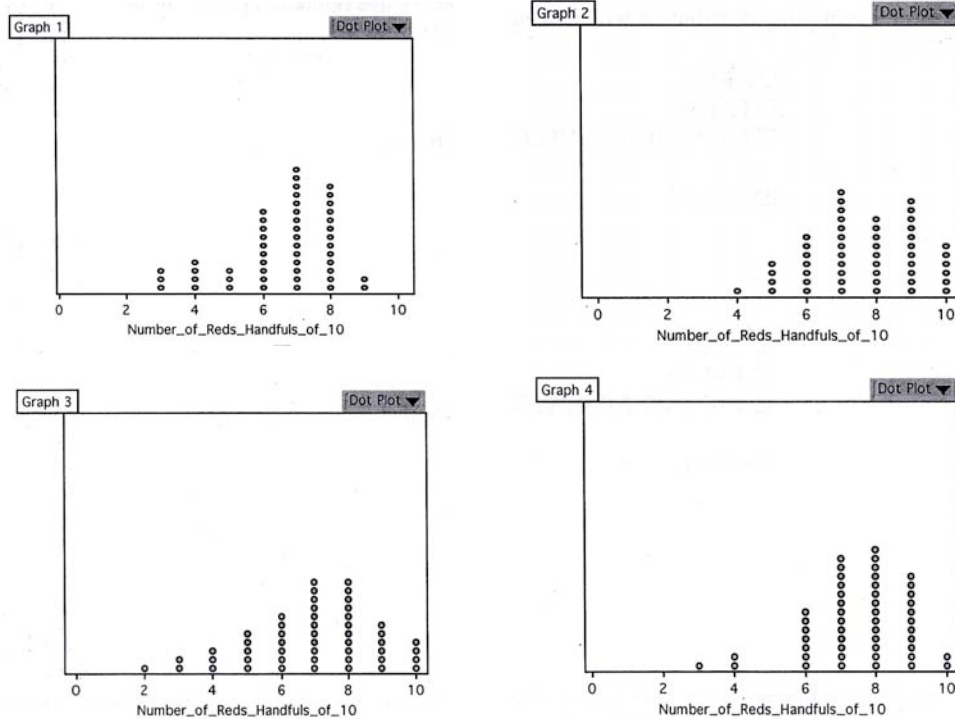
Later in the interviews the question was rephrased and TAs were asked to rate the

graphs from most likely to least likely. This task is essentially about detecting

suggested fraud by identifying classes of graphs that are more likely to occur and

classes of graphs that are less likely to occur. Graphs 1 and 3 were manufactured

('fake') and Graphs 2 and 4 were generated by computer simulation ('real')

(Shaughnessy et al., 2004a&b). Graph 1 was designed with an appropriate range, but

shifted to the left. Thus, Graph 1 has an unusually high number of outcomes at four

and below and too few at nine and ten. Graph 3 was designed to have a smooth

distribution in frequency and a range that is unusually wide.

A Model of Reasoning about the Real/Fake Task

There were three primary reasoning paths which TAs employed to justify their

real/fake decisions: (1) attention to a single attribute – shape, center, outlier, or spread;

(2) attention to two or more attributes, using an informal distributional argument; and,

(3) attention to the underlying probability distribution – left or right end of the

distribution. Figure 19 provides a visual model for this conceptual framework.

Figure 19: Real/Fake Task Conceptual Reasoning Framework
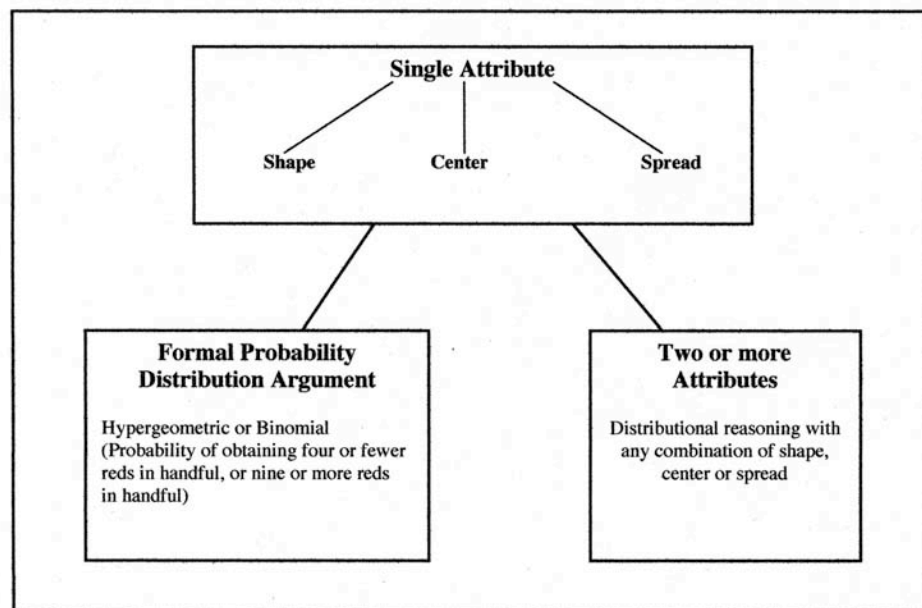
TAs could focus on a single attribute (shape, center, or spread) of the distribution.

If the focus is on shape, then attention could be directed toward the variability in

frequencies on the graphs of each empirical sampling distribution (i.e., the varying

heights on the graph of the empirical sampling distribution as one moves from left to right). If the focus is on measures of center, then attention could be directed toward the mode, median, mean, or any combination thereof. Table 7 shows the mode, median, and mean for each of the four graphs.

Table 7: Measures of Center for Real/Fake Graphs

Measures of Center	Graph 1	Graph 2	Graph 3	Graph 4
Mode	7	7	7 & 8	8
Median	7	8	7	8
Mean	6.54	7.62	6.92	7.5

The mode and the median are not particularly helpful measures of center for detecting which graphs are fabricated, in that the mode and median for each graph is either a seven or an eight. The mean could be helpful because one can make a comparison of the mean of each graph against the expected mean of 7.5, based on the population proportion. If the focus is on spread, then attention could be directed toward the range, interquartile range (IQR), variance, standard deviation, or the left and/or right ends of the distribution. Table 8 shows the range, IQR, and standard deviation for each of the experimental graphs.

Table 8: Measures of Spread for Real/Fake Graphs

Measure of Spread	Graph 1	Graph 2	Graph 3	Graph 4	Theoretical Model
Range	6	6	8	7	6
IQR	2	2	2	2	1
Standard Deviation	1.54	1.56	1.93	1.46	1.36

The different measures of spread, range, IQR, and standard deviation may not be as helpful for detecting the real graphs versus the fake graphs. The overall range of the data set is not large, and the standard deviation would be more time consuming to compute during a survey or interview situation. However, a focus on the outlying ends of the distribution in this situation may be helpful in distinguishing the real graphs from the fake. An estimation of the likelihood of handfuls containing few (0, 1, 2, or 3) red candies, or the likelihood of handfuls containing many (9 or 10) red candies, could be of use in distinguishing the real graphs from the fake graphs. Reasoning with two or more attributes of the distribution constitutes an informal distribution argument.

In a formal probability distribution argument, the TA relates the situation back to the theoretical probability model. The context for the Real/Fake Task is the same as the Prediction Task, so either the hypergeometric or binomial could serve as the underlying probability structure. In using the underlying probability structure as a tool for determining the real graphs from the fake graphs, a number of probabilities could be computed: (a) the probability of getting six or more handfuls containing four or fewer reds[20] (Graphs 1 & 3); (b) the probability of getting 17 handfuls containing nine or more reds (Graph 2); or, (c) the probability of getting two handfuls containing nine or more reds (Graph 1). Using a binomial probability distribution for 50 trials of the experiment, the probability of obtaining four or fewer reds is approximately

[20] All of the TAs from the survey using a formal probability distribution argument discussed the extremely low probability of getting too many handfuls with four or fewer reds. I discuss this in more detail in the next section.

0.019193288. Thus, the probability of six or more trials resulting in four or fewer reds is the summation of:

$$\binom{50}{6}(0.019193288)^6 (1-0.019193288)^{44} \approx 0.00033862 \qquad (1)$$

$$\binom{50}{7}(0.019193288)^7 (1-0.019193288)^{43} \approx 0.000041652 \qquad (2)$$

$$\binom{50}{8}(0.019193288)^8 (1-0.019193288)^{42} \approx 0.000004381 \qquad (3)$$

$$\binom{50}{9}(0.019193288)^9 (1-0.019193288)^{41} \approx 0.0000004001 \qquad (4)$$

$$\binom{50}{10}(0.019193288)^{10} (1-0.019193288)^{40} \approx 0.0000000321 \qquad (5)$$

Summing (1) through (5) is approximately 0.0003851. Investigating the left end of the distribution in this manner demonstrates the unlikelihood of graphs like Graph 1 and Graph 3. The probability of obtaining nine or more reds is approximately 0.2426. Thus, the probability of 17 out of 50 trials resulting in nine or more reds is:

$$\binom{50}{17}(0.2426)^{17} (1-0.2426)^{33} \approx 0.0358261$$

Comparing this probability with the previous probability for obtaining six or more handfuls with four or fewer red provides further evidence that types of graphs like Graphs 1 and 3 are less likely. That is, the probabilities calculated here provide evidence that it is less likely to pull out a few (six or more) handfuls containing four or fewer reds than it is to pull out many (17) handfuls containing nine or more reds.

Connecting Theoretical Expectations to Experimental Data

In addition to the manner in which a TA reasons about the graph (single attribute, informal distribution, or formal probability distribution), a TA either expects or does not expect the experimental distributions to match up with the theoretical model, and their own image of the theoretical model could be different than the actual model. The TA's expectations about experimental data and their own image of the theoretical model play a stronger role in the TA's decisions about the real/fake graphs compared to the Prediction Task. In the Prediction Task, TAs could simply provide the theoretical predictions calculated from the hypergeometric or binomial models, or TAs could provide a prediction based on what they think might actually occur during the experiment. The Prediction Task is open-ended in that respect. However, in the Real/Fake Task, TAs are forced to make decisions about experimental data (real and non-real); thus, their expectations of experimental data and their concept image (in the sense of Tall & Vinner, 1981) of the theoretical model have a greater impact on their choices.

Having presented a conceptual analysis and framework for the Prediction and Real/Fake Tasks, I set forth with a discussion about how the TAs reasoned in this context, grounded in the conceptual analyses. I begin with a discussion of how TAs reasoned about the Prediction Task and move to how they reasoned about the Real/Fake Task. I focus on the ways in which TAs did or did not use their knowledge of the underlying probability structure to resolve tensions when examining experimental data. TAs' expectations of experimental data and their own concept

image of the theoretical model appear to be particularly germane to their ability to resolve tensions between theoretical models and experimental data. In particular, TAs' conceptions of variability in this context appeared to be a key source of tension that they could not always resolve in a statistically coherent manner. I begin with a discussion of the survey results, followed by more robust analysis from four interview participants.

4.1.3 TA Thinking and Reasoning about the Prediction Task

Of the TAs who participated in my survey, 79.4% (N=68) used either an informal or formal distributional argument to support their prediction on the Prediction Task (see Table 9).

Table 9: Justifications for the Prediction Task

Formal Probability Distribution	Informal Probability Distribution	Center	Other (Not able to code)	Total
45 (66.2%)	9 (13.2%)	6 (9%)	8 (11.8%)	68 (100%)

Most of these TAs used either a binomial or hypergeometric probability distribution argument. A few TAs argued informally about the ratio of reds to yellows in the jar coupled with comments such as "distributes evenly around the center", or "goes down on either side from the center". The responses from the survey provide some indication that these TAs are comfortable reasoning distributionally and can determine underlying probability structures in this context. However, many of the TAs in this study appeared to be situated at the left end of the continuum, in that they provided predictions where the spread is more consistent with an inexact mapping to the

117

theoretical. Table 10 shows that approximately 54% of the TAs in this study provided predictions that matched my four established criteria (whole number, center, shape, and range – recall Figure 17). About 6% of TAs deviated solely from the whole number criteria. The remaining TAs provided predictions that disagreed with at least one or more criteria related to the distribution, situating them closer to the left end of the continuum with predictions that deviate more radically from the theoretical model in at least one attribute.

Table 10: Prediction Task - Ranking TAs' Predictions

Number of Matches with Four Criteria					
0	1	2	3	4	Total
7 (10.3%)	1 (1.5%)	9 (13.2%)	14 (20.6%)	37 (54.4%)	68 (100%)

Approximately 24% of TAs in this study provided predictions that deviated from the range criteria. These TAs provided ranges eight or more units wide, indicating that these TAs may expect a wider range in experimental data than the theoretical model suggests is likely. That is, perhaps these TAs did not have a mental scheme for the concept of bounded variation – the balance between sample variability and sample representativeness. This finding is consistent with studies involving K-12 students (Reading & Shaughnessy, 2004; Rubin et al., 1991; Saldanha & Thompson, 2003).

Table 11 provides the reason codes and number of matches to my four criteria for four of the TAs I interviewed. On the survey, Sandy and Andy provided formal probability distribution arguments, Amanda provided an informal distributional argument, and Joe provided a proportional argument. During the interviews it became

clear that Amanda also used a formal probability distribution argument and Joe an informal probability distribution argument. Only Sandy's and Joe's predictions matched all four of my criteria.

Table 11: Four Case Studies – Type of Reasoning Employed and Number of Matches to Prediction Task Criteria

	Amanda	Sandy	Joe	Andy
Reasoning Code based on the survey	Informal Distributional	Formal Distributional	Proportional	Formal Distributional
Reasoning Code based on the interview	Formal Distributional	Formal Distributional	Informal Distributional	Formal Distributional
Whole Number Criteria	✔	✔	✔	
Center Criteria	✔	✔	✔	✔
Shape Criteria	✔	✔	✔	✔
Range Criteria		✔	✔	✔

Andy's prediction did not map to my whole number criteria, and Amanda's prediction did not map to my range criteria because her range is greater than eight units (2 red candies to 10 red candies). Table 12 shows the predictions provided by Amanda, Sandy, Joe and Andy.

Table 12: Interview Participants' Predictions for Prediction Task

Number of Red Candies in Handfuls of 10	Predictions			
	Amanda	Sandy[21]	Joe	Andy
0 red	0	0	0	0
1 red	0	0	0	0
2 red	1	0	0	0
3 red	1	0	1	0.1
4 red	1	1	1	0.8
5 red	3	3	3	2.9
6 red	7	8	9	7.3
7 red	12	13	12	12
8 red	12	14	12	14
9 red	9	9	9	9.4
10 red	3	3	3	2.8

Andy used the hypergeometric model to calculate his predictions. He did not round to the nearest whole number despite the candy jar context, suggesting a purely theoretical, rather than an experimental prediction. Sandy also used the hypergeometric model to calculate her prediction, but she rounded to the nearest whole number because of the context. The prediction provided by Sandy is typical of the predictions TAs provided in the survey – approximately 54% of TAs on the survey provided literally the same prediction. Joe reasoned with a proportional argument on the survey, but his prediction is similar to Sandy's. The main distinction between Joe's prediction and Sandy's is that Joe made his distribution symmetric around the population proportion. Finally, Amanda's prediction is also similar to Sandy's, but it

---

[21] Sandy's prediction sums to 51; this is her error. It appears that she may have rounded up from 7 to 8 handfuls containing six red candies after she applied the hypergeometric model.

spreads further out into the left tail. Amanda predicts one occurrence where there will be two red candies and one occurrence where there will be three red candies.

During the first interview I followed up with these four TAs to get a better sense of how they thought about the Prediction Task. From the interview it was clear that Amanda, Sandy, and Andy could relate the Prediction Task to the binomial or hypergeometric probability distribution. Amanda, Sandy, and Andy were aware of the conditions on each of these distributions, and that the population of candies was large enough to allow for a binomial approximation. I begin with a discussion about Sandy's and Andy's approaches to the Prediction Task. I will follow with a discussion of Amanda's approach and conclude with Joe's approach.

Sandy's and Andy's Approach: Hypergeometric Distribution

Sandy and Andy both used the hypergeometric distribution, and their predictions reflect the outcomes calculated from that distribution. The Prediction Task appeared to be a relatively straightforward application of the hypergeometric for Andy and Sandy. They both knew the probability distribution formula and the conditions that needed to be met before it could be applied. When I questioned them about using the binomial model instead of the hypergeometric model they both recognized that either model is appropriate in this context because of the large population ($N = 1000$) and the small sample size ($n = 10$).

Amanda's Approach: Binomial Distribution

Amanda's survey response was coded as informal probability distribution, but there was strong evidence from our conversation during the interview that she did in fact use the binomial to help her make her predictions.

Interview 1: Prediction Task:

> **Amanda**: Well you're looking at 750 red and 250 yellow, so on average I'm expecting between 7 and 8 reds per handful, based on the percentage of red and yellows that are in the jar of 1000 candies. Given that the ratio is pretty skewed in the red direction, I'm most likely going to be pulling reds out of this jar given that there are 750 to the 250 yellow. I did not think that it was reasonable that on any one of my draws that I'm going to have no red candies or even 1 red candy. I'm thinking every time I reach in I'm going to have a red candy. I think the probability of getting no red candies is excruciatingly small. So I think that out of the 50 it's not going to happen. I know that on average each handful of 10 are going to have about 7 or 8. So I lumped most of my 50 into 7 and 8 [*red candies*] and then I tried to disperse it around the 7 and 8 appropriately. And I did some calculations.
> **Interviewer**: So what kind of calculations?
> **Amanda**: I did some binomial calculations.
> **Interviewer**: Alright, so why binomial?
> **Amanda**: Because I have essentially two outcomes. I have red and yellow. I have the probability of red occurring and the probability of yellow occurring, .75 versus .25. And I'm going to reach in and grab 10, and I want to know what the probability is of getting a certain amount of them red. And then I can use that to examine all 50.

The first highlighted utterances in the previous excerpt provides some explicit evidence that Amanda recognized the ratio of red to yellows in the jar, and that there would be some amount of spread around that center. Amanda recognized the ratio of red to yellow and how the distribution was skewed toward red candies. This knowledge seemed to provide Amanda with a sense for how unlikely it was for the experiment to produce handfuls containing zero or one red candy. Amanda was able to discuss multiple aspects of the distribution. She used her knowledge of the ratio of red

122

to yellow and her understanding of variability from trial to trial as a means for making her predictions around seven and eight red candies. In fact, the second and third highlighted utterances suggest that she used a binomial probability distribution to help her with her calculations. In the last highlighted utterance she provides justification for using a binomial probability distribution by discussing two of the four assumptions of the binomial probability distribution – each trial results in one of two outcomes and the probability for each of those outcomes is constant from trial to trial.

Joe's Approach: Informal Distribution

On the survey, Joe's justification for his predictions focused primarily on the ratio of red to yellow candies in the jar. This remained his primary focus during the interview as well, yet Joe did have a sense that there would be some variation from handful to handful and he had a particular image of the shape of the distribution. Thus, Joe appeared to reason overall with an informal distribution argument.

Interview 1: Prediction Task:

> **Joe**: So I thought because 3/4 of them are red and 1/4 of them are yellow in the jar, the most likely, if I grab 10 of them, is 7.5, or between 7 and 8 [*red candies*]. So if I kept doing this I expected to see between 7 and 8 equally likely, and those the most likely to get. So then I just made up some numbers from there. I thought 9 and 6 [*red candies in a handful of 10*] were right next to them [*7 and 8 red candies in a handful of 10*] so I gave it 9. And these are just guesses. I didn't do the math.
> **Interviewer**: So it looks like you kind of made it symmetrical around that 7.5?
> **Joe**: Yeah, and then a bit of a tail out here [*to the left*]. I mean the chance of getting 0 red candies is phenomenally low so I just didn't expect to see any one of those in 50 trials.
> **Interviewer**: You didn't do any calculations. You just did this about reasoning the number of reds to yellows in the jar….Some folks were using

different distributions, like the binomial and the hypergeometric when they were doing this problem. So would you use one of those? Do you think one of those are appropriate here?

Joe: [*Laughs*] To tell you the truth, I don't know enough about them. I don't know what those distributions are or do really. I mean it felt really disconnected from the rest of the material like this is a formula we are presenting. Now it's plug and chug and go.

There are three main points of interest in the previous excerpt. First, in the first highlighted utterance, Joe was primarily reasoning by the ratio of red to yellow, noting that ¾ of the jar contains red candies. At the same time, Joe had some strong intuitions for spread around that center value. Although Joe did not directly articulate the idea of variability or spread in this excerpt, he attended to that attribute of distribution, particularly when he indicated how unlikely it was to grab 10 candies without a single red in the handful.

Second, implicit in Joe's discussion is the attribute of shape. Joe's attention to shape can be seen in his utterances about having an equal number of occurrences on seven and eight red candies followed by an equal drop to the left of seven red candies and to the right of eight red candies. In fact, it is not by chance that Joe made his distribution symmetric about seven and eight red candies. It appears that Joe believed the shape of the distribution would be symmetric. Joe's image of the shape of the distribution is a point that deserves further attention because it is unclear from Joe's utterances if he believed that graphs of experimental data have the same smooth shape as the underlying theoretical model. Joe's prediction might be a projection of his image of the theoretical model, the experimental data or both.

Finally, Joe did not make a connection to an underlying probability structure in the Prediction Task context. When I asked Joe about using a binomial or hypergeometric distribution to think about the Prediction Task, the last highlighted utterance suggests that he was not familiar enough with these distributions to comment on how they might relate to the Prediction Task. On an intuitive level, Joe appears to have a strong sense of shape, center and spread for a distribution of data, but he has not connected these attributes to a formal probability structure.

4.1.4 TAs' Thinking and Reasoning about the Real/Fake Task

There was an unusual shift in TAs' reasoning from the Prediction Task to the Real/Fake Task that provided the first indications that these TAs might experience tension when making decisions about experimental data. Given that most TAs in this study used a formal or informal probability distribution argument to justify their responses to the Prediction Task, I expected a similar justification for their decisions on the Real/Fake Task. Yet, TAs tended to justify their choices on the Real/Fake Task based on a single attribute of the distribution. Only 17.7% (N=68) of TAs used a formal or informal distributional argument to justify their real/fake identifications, compared to 79.4% on the Prediction Task (see Table 13).

Table 13: Real/Fake Task Responses

Number Correct Identifications	Single Attribute (Center/ Spread/ Shape)	Single Attribute (Shape Only)	Informal Probability Distribution	Formal Probability Distribution	Other	Total
	Primary Reasoning for the Real/Fake Task					
0	0	2 (2.9%)	0	0	1 (1.5%)	**3 (4.4%)**
1	1 (1.5%)	3 (4.4%)	0	0	0	**4 (5.9%)**
2	4 (5.9%)	15 (22.1%)	0	0	5 (7.5%)	**24 (35%)**
3	1 (1.5%)	16 (23.5%)	0	1 (1.5%)	0	**18 (26%)**
4	8 (11.8%)	0	8 (11.8%)	3 (4.4%)	0	**19 (28%)**
Total	**14 (20.6%)**	**36 (53%)**	**8 (11.8%)**	**4 (5.9%)**	**6 (8.8%)**	**68 (100%)**

Table 13 shows the distribution of the types of reasoning TAs employed, crossed with their number of correct identifications. There are two columns for Single Attribute reasoning. The first column groups together TAs that used the same single attribute (center or spread) of the distribution as their primary reasoning for all real/fake graphs with those that used a single attribute (center, shape, or spread), but that attribute may have been different for different graphs (e.g., Shape for Graph 3 and Spread for Graph 1). Many, TAs based their decisions in the Real/Fake Task on the attribute of shape. In fact, 53% of TAs used a shape argument only to justify their real/fake identifications and another 10% used a shape argument to justify their

126

identification for Graph 3 and a different single attribute for the other graphs (e.g., center). This is an interesting result in light of the fact that 0% used a shape argument on the Prediction Task. Perhaps TAs' focus on shape is a result of the graphical display of the experimental data.

Recall that Graphs 1 and 3 were manufactured ('fake'), while Graphs 2 and 4 were generated by computer simulation ('real'). Table 14 shows the distribution for the number of matches in TAs' identifications.

Table 14: Correct Identifications in the Real/Fake Task

0 matches	1 matches	2 matches	3 matches	4 matches	Total
3 (4.4%)	4 (5.9%)	24 (35.3%)	18 (26.5%)	19 (27.9%)	68 (100%)

Approximately 46% of the TAs taking this survey provided 0, 1, or 2 matches. One's chances of guessing and getting two matches are 50%. This means that close to half the TAs performed no better than they would have if they had simply guessed which graphs were real and which were made-up. In addition, 16 of the 18 TAs who got three matches based their decision solely on shape. In general, these TAs marked Graph 3 as made-up since its shape was 'too smooth' and the other three as plausible because the 'bumps', 'ups & downs', and general 'unevenness' are more likely to happen in a real sampling situation. To focus on one aspect of the distribution in these graphs is not sufficient for making a determination about whether or not a graph of experimental data is likely to be fraudulent. Those TAs who used some form of distributional argument made correct identifications on all four graphs (with one exception getting 3 matches). It seems likely that in order to successfully identify

127

unlikely classes of graphs requires the ability to coordinate multiple attributes of a distribution and the ability to maintain a sense of bounded variability for the outcomes of the experiment. Interestingly, the TAs' main criteria for accepting a graph as real appeared to be based on unevenness in the frequency of the graph, which most TAs expected because of "natural variation in the sampling process". That is, the basis of TAs' attention to variability appeared to be grounded in the shape of the graphs. In particular, TAs appeared to focus on the variability in the vertical heights of the distribution, rather than on variability in a statistical sense.

Table 15 provides a closer examination of four of the TAs I interviewed. Table 10 shows the number of matches in the TA's prediction with my four criteria for the Prediction Task, the reasoning code assigned to their justification on the Prediction Task, the number of correct identifications on the Real/Fake Task, and the reasoning code assigned to their justification on the Real/Fake Task. Amanda, Sandy, Joe and Andy's responses to the Prediction Task and the Real/Fake Task were representative of the majority of TAs' responses on the survey.

Table 15: Comparison of responses from Prediction Task to Real/Fake Task

TA	Prediction Task (Number of matches on 4 criteria)	Justification for Prediction	Number of Matches on Real/Fake	Justification for Real/Fake
Amanda	Matched 3 criteria	Formal Distributional	2	Shape
Sandy	Matched 4 criteria	Formal Distributional	3	Shape
Joe	Matched 4 criteria	Informal Distributional	2	Shape
Andy	Matched 3 criteria	Formal Distributional	2	Shape

Each of these TAs used formal or informal distributional arguments to justify their predictions, yet when they examined the real/fake graphs they made their identifications based primarily on shape. These TAs, like many of the TAs in the survey, used language like Graph 3 is "too perfect" or "too smooth" to be real, whereas Graph 1 (or one of the other graphs) has "ups & downs" or "natural variability", so it is likely to be real. These types of shape-oriented responses are similar to the shape oriented responses of middle and secondary school students on the Real/Fake Task (Shaughnessy et al., 2004b). From reading the survey responses alone, I could not be sure whether when TAs used the phrase "natural variability" they were referring to the variability in the frequencies for each graph (i.e., changes in the vertical heights), or if they were referring to variability in terms of variance. Evidence from the interview data suggests that such references were describing changes in frequency because during the interviews each TA appeared to describe natural variability in terms of changes in frequency from one outcome to the next; indeed, variance never entered the conversation. Torok and Watson (2000) found similar results in their work with K-12 students.

In the subsections that follow, I discuss TAs' thinking about the Real/Fake Task. In particular, I discuss how these TAs primarily focused on shape, and how that focus influenced their decision-making process in light of their image of the theoretical model and how closely they expected experimental data to match the theoretical model. Second, I discuss the tensions TAs experienced as they attempted to make decisions based on experimental data.

Focus on Shape: "Graph 3 is too perfect"

Each of the TAs I interviewed primarily focused on the attribute of shape in order to distinguish a fraudulent graph from an actual computer generated graph. Amanda, Sandy, Andy and Joe identified Graph 3 as made-up because its shape is "too perfect" to be real. In the excerpts that follow, I chronicle what each of these TAs had in mind by the phrase "too perfect". First, I begin with an exchange that provides some insight into what Amanda meant by "too perfect". The highlighted utterances suggest that the steady, smooth increase from the left of the graph to the center, followed by the steady, smooth decrease from the center to the right of the graph, is the type of shape Amanda expects to see in theoretical models.

Interview 1: Real/Fake Task:

> **Interviewer:** When you say Graph 3 is too perfect, what do you mean by too perfect?
> **Amanda:** I expect in a real graph that they're not going to be in order 0, 1, 2, 3, 4 etcetera. Some are going to be higher than the one next to them and some are going to be lower.
> **Interviewer:** So the even steps up in frequency?
> **Amanda:** Yes, it goes up very smoothly. It doesn't have any dips in terms of 0 [*red candies*] up to the 7, 8 [*red candies*] and then from the 7, 8 [*red candies*] it's decreasing back down. And that just strikes me as too perfect. Compare that to Graph 1 where we increase [*in frequency*] from 3 to 4 [*red candies*] but then we decrease from 4 to 5 [*red candies*]. So it increases too smoothly and too evenly. I just don't buy this. It's not monotonic. Monotonic is not the right word because it dips back down, but on either side of 7 and 8 it's monotonic. It's just increasing and then just decreasing. Which if I could repeat the experiment to infinity, I might expect to happen.

Also notice that Amanda compares the smoothness of Graph 3 to the unevenness in Graph 1. In the next exchange, it appears that Amanda identified Graph 1 as real, in

large part because of the dips in frequency, something she seemed to expect in

experimental data.

Interview 1: Real/Fake Task:

Interviewer: So is this [*Graph 3*] also your image of what the ideal graph would look like, sort of in line with your own predictions?
Amanda: Yes.
Interviewer: So, it's almost like you're telling me, and I don't want to put words into your mouth, but Graph number 3 is sort of the way the theoretical graph would look to you?
Amanda: Yes. Yes. Yes it is.
Interviewer: And experimentally this is…
Amanda: Not going to happen [*laughs*].
Interviewer: [*Laughs*]. So Graphs 1 and 4 were real because they have these things that do occur experimentally?
Amanda: Yeah, all these little quirks. We have in Graph 1, more piled on 4 [*red candies*] than we do on 5 [*red candies*]. Graph 4 we don't have anything at 5 [*red candies*], these are things that occur in an actual testing situation.

In this excerpt, Amanda indicates that experimental graphs will have "quirks" like

gaps or dips up and down in the frequency as we move from left to right. Thus, her

main criteria for identifying the real versus the fake graphs appears to be based on

'quirky' shapes that she expects to get in experimental situations.

Sandy and Andy also indicated that Graph 3 was "too perfect" to be real, and

Graph 3, at least in terms of shape, seemed to match their image of the theoretical

distribution.

Interview 1: Real/Fake Task:

Interviewer: What do you mean by too perfect? Because that's something a lot of TAs said about this one [*pointing to Graph 3*]?
Sandy: Because, too perfect is like what you expect it to have this shape [*drawing a left skewed curve over the bar graph*].

Interviewer: And you think of the four of these [*graphs*] this one [*Graph 3*] is closest to the theoretical distribution?
Sandy: Because it has look [*makes a sketch of a left skewed distribution over the Graph 3*] this kind of shape, you see - close to 0 and then mounds up and then decreases.

Interview 1: Real/Fake Task:

Andy:… It's not ideal. I expect it to have [*makes a gesture – draws a distribution curve in the air with lots of vertical ups and downs*].… It's going to have like defects, all sorts of holes, funny anomalies. Like you know, like maybe a case out here [*pointing to the 1, 2, and 3 red candies range on graph 1*] or maybe like this one [*pointing to Graph 1 at the 3, 4, and 5 red candies spots*], like this divot [*where Graph 1 dips down at 5*]. Where if the student was cheating, you know if that's the case, then they're going to do something more like this [*pointing to Graph 3*]. Where you are going to have a nice up curve. I mean you could almost draw a nice curve over this.
Interviewer: So is this what you mean by too perfect? One of my questions for you is that you wrote that the last two graphs seemed too perfect.
Andy: That's what I mean, too close to the ideal when it's such a small sample.

In both of these excerpts, Sandy and Andy made similar comments as those made by Amanda about the qualities of the 'ideal' graph. It seems that Amanda, Sandy, and Andy expect theoretical models to have a nice smooth shape, but they expect experimental data to have more variability in the frequency. That is, these TAs appear to expect increases and decreases in frequency, rather than a constant increase followed by a constant decrease. Also, like Amanda, Andy provided some indication of his image of experimental data. Andy's utterance about graphs of experimental data having "defects", "holes" or other anomalies appeared to be in reference to changes in frequency. There was no evidence that Amanda, Sandy, and Andy attended to other attributes of the distribution during our conversations.

Like Amanda, Sandy, and Andy, Joe also focused on the shape of Graph 3. Joe suggested Graph 3 was made-up because it was "too normal looking".

Interview 1: Real/Fake Task:

> **Interviewer:** You said that Graphs 1 and 4 were real and Graphs 2 and 3 were made-up. Tell me a little bit more about what you mean by Graph 3 looks way too normal?
> **Joe:** Yeah, I use the normal to mean regular because the normal distribution has some characteristics…. I would expect over time if I was doing these 50 trials of 10 and I did a bunch of them, I would expect those averages to look something like this [**pointing to Graph 3**].
> **Interviewer:** Okay, so this is your image of the theoretical distribution for this problem?
> **Joe:** Fairly close, I mean getting two reds [**points to the frequency of 2 in the 2 red candies slot**] I'm not sure about the numbers, how those worked out. It seemed fairly unlikely, but I didn't run any numbers on this.

The first highlighted utterance suggests that Graph 3 is Joe's image of the theoretical model, at least in shape. In this respect, Joe's initial reactions to the Real/Fake Task are similar to the reactions of Amanda, Sandy and Andy. Yet, there is some evidence that Joe also attended to the tail of the distribution in Graph 3 when he mentioned the likelihood of pulling out a handful containing two reds. Joe's attention to the left tail is important because it provides explicit evidence that he looked at Graph 3 as the ideal graph more strictly in terms of shape rather than any other aspects of the distribution. He recognized that some of the outcomes, in particular the number of handfuls containing two red candies, might, in fact, be quite unlikely. He could not say for sure how likely certain outcomes were, however, because he did not have the computational tools yet.

Amanda's, Sandy's, Andy's and Joe's articulations of the shape of Graph 3 during

the interviews suggest that the shape of Graph 3 fit their image of the shape of the

theoretical model. In addition, their images of the shape of the theoretical model map

to the actual shape of the theoretical model – a left-skewed distribution. Also, from

these excerpts it appears that each of these four TAs did not expect the shape of

experimental data to *exactly* match with the shape of the theoretical model. I did not

find any evidence in the interview excerpts to suggest that these TAs focused on and

incorporated other attributes of the distribution, such as center and spread, in making

their determination.

Influence of Shape on TAs' Real/Fake Decisions

Formulating the different components of TA reasoning within the previously

discussed conceptual framework and continuum scale for this task, Figure 20 provides

a mapping of reasoning that appeared to lead each of these TAs to reject Graph 3.

Figure 20: Influence of Shape in Real/Fake Task
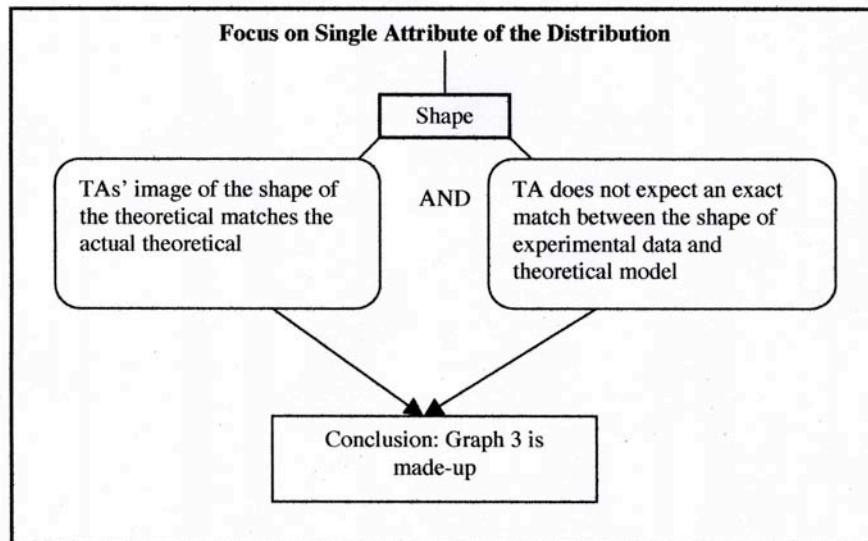
It appears that each of these TAs concluded Graph 3 was 'fake' because their image of the shape of the theoretical graph matches closely with the actual shape of the theoretical model, and their expectations for the shape of experimental data fell somewhere on the continuum from expecting large deviations to small deviations. That is, they did not expect the shape of experimental data to *exactly* match that of the theoretical model. Yet, a focus solely on shape was not especially helpful for making determinations about the other three graphs. Table 16 shows the graphs that each TA deemed 'fake' and 'real'. It appears that a focus on shape alone is likely to create agreement only for determining that Graph 3 is made-up.

Table 16: Real/Fake Identifications

TA	Graphs identified as real	Graphs identified as fake
Amanda	1 & 4	2 & 3
Sandy	1, 2, & 4	3
Andy	1 & 2	3 & 4
Joe	1 & 4	2 & 3

Amanda believed that Graphs 1 and 4 were real, while Graphs 2 and 3 were made-up, because Graphs 1 and 4 had more "quirks".

Interview 1: Real/Fake Task:

Interviewer: So Graph 1 and 4 were real because they have these things that do occur experimentally?
Amanda: Yeah, all these little quirks. We have in Graph 1, more piled on 4 [*red candies*] than we do on 5 [*red candies*]. Graph 4 we don't have anything at 5. These are things that occur in an actual testing situation.

Whereas Amanda believed the gap at five red candies on Graph 4 was something likely to occur in actual experiments, Andy believed that the divot at the eight red candies spot on Graph 2 was the kind of defect one could expect in experimental data.

Interview 1: Real/Fake Task:

Interviewer: So when you talk about defects is that, like when you talked about Graph 2 as being biased some how…
Andy: Yeah, it's biased. Look at how it's wandered away from the mean [*pointing to the low number of values on the 8 reds spot*].

Further, in contrast to Amanda, Andy believed that since the divot at the five red candies spot on Graph 4 was the only 'defect' in Graph 4, it could be made-up.

Interview 1: Real/Fake Task:

Interviewer: So you said this one [*Graph 4*] was fake for the same reason?
Andy: Okay, let me see. This one. Oh, I don't know. I would say this one is a little more marginal, but if you forced me to put it in a category I'd say

136

maybe it's fake, just because you can draw the curve over top of it. And if you're, and it depends on how skilled the fraud person is. They may say, 'I know you can't get a perfect one so I'll put a defect in'. So there's my defect [*points to the gap at 5 on Graph 4*]. You know. You didn't get any at 5. It's a defect from the ideal and now the fraud is undetectable.

Andy believed that Graphs 3 and 4 were made-up because the shape of Graph 3 was clearly "too perfect" and the shape of Graph 4 was close to the ideal except for one "defect". However, Andy did not feel as confident in identifying Graph 4 as a 'fake' (notice in the first highlighted utterance in the preceding excerpt).

Joe used a similar argument for why Graph 2 could be 'fake'. Except that the defect for Joe was the divot at eight red candies on Graph 2, rather than the divot at five red candies on Graph 4.

Interview 1: Real/Fake Task:

Interviewer: And you said Graph 2 was fake for kind of a similar reason?
Joe: Yeah, so you know. It looked a little too normal. It had this divot here [*Graph 2 at the 8 red candies spot*]. And again, this is the psychology of the student than the actual distribution. Because I would think, you know if the student was faking the data, not actually doing the trials, but just faking it. This [*Graph 2*] would be a little more clever than this [*Graph 3*] in that the student would say 'okay, it's not going to be exactly that so I'll mess it up a little bit'. And I thought that, really I wanted to pick two that were fake and two that weren't, is part of what motivated me.

Like Andy, Joe thought that a clever student would throw a requisite defect into his/her graph. Yet, they both gravitated to different defects and suggested that except for the small defect, the graphs followed a nice smooth shape, like the theoretical, and were thus likely to be made-up.

On her survey, Sandy marked Graphs 1, 2, and 4 as real because they all had shapes that deviated in some fashion from the "perfect" shape of Graph 3.

137

Interview 1: Real/Fake Task:

> **Sandy:** So here you said the classes conducted the experiment. Since the class conducted the experiment, this is what I believe. So again this is something that I believe because I didn't do this kind of stuff, you know. But I believe that in nature when you try things they are supposed to be more randomly than the theoretical. So, these seemed like something reasonable, for me, to happen because they have this weird shape [*points to Graphs 1, 2 and 4*]. But who knows? Maybe you are lucky and get here [*points at Graph 3*].

Like Andy and Joe, Sandy seemed to be expressing some uncertainty in her response. Each of these TAs seemed convinced that Graph 3 was fraudulent because it was too smooth, but the other graphs had some deviations in frequency that made the shape differ from the theoretical model.

These deviations in frequency seemed to create a source of tension for the TAs as they set out to determine which of the other graphs could possibly be fraudulent. The source of the tension is over how "different" the graphs of the experimental data can be from the theoretical model without being considered unusual. For a statistician, this tension is resolved through robust knowledge of the concept of bounded variability. That is, a deep understanding of the balance between sample representativeness and sample variability. The TAs in this study resolved this tension by looking at the 'unevenness' in the frequencies of each graph. Focusing on the 'unevenness' in frequencies is helpful for determining that Graph 3 is 'fake', but this method becomes problematic for making determinations about the other graphs, especially when there is no attention on the center and spread of the graph. For instance, Graph 1 has more variation in frequency, but its center is unusually low and there are too many outcomes

138

with four or fewer reds to be plausible. Graphs 2 and 4 have less variation in frequency, making it difficult to eliminate or retain these graphs on that basis alone.

In addition, these TAs seemed to rely on a subjective determination of how closely the shape fit the theoretical shape in order to justify their responses. In particular, Amanda, Andy, and Joe resolved this tension by choosing a graph (or graphs) that they *believed* either modeled the type of defects one could expect in experimental data and labeled that graph as 'real', or they chose a graph that they believed resembled the theoretical model with the exception of one small defect and labeled that graph as 'fake'. Sandy resolved this tension by choosing to consider all the remaining graphs as 'real' because they all contained some variation in their shape as compared to the theoretical model.

Follow-up Questions: Alternative Wording of the Real/Fake Task

After my initial follow-up questions about the Real/Fake Task in the interviews, I changed the wording of the problem and asked TAs to rate the graphs from most likely to least likely. I wanted to know if the new phrasing of the problem would change how TAs perceived the task. In general, it did not. These TAs continued to think about the task in terms of what they expected to see in experimental data, mostly from a shape point of view. However, the extended discussion around the likelihood of the different graphs in the Real/Fake Task provided additional insight into TAs' thinking about experimental data and how they resolve their tension between experimental data and the theoretical model.

In the subsections that follow, I discuss the reasoning of Amanda, Sandy, Andy, and Joe in light of the conceptual framework, continuum of expectations for experimental data, and TAs' own image of the theoretical model discussed earlier in the chapter. I begin with a discussion of Amanda, followed by Sandy, Andy and Joe.

Amanda

When I asked Amanda to rate the graphs from most likely to least likely she responded that Graph 3 was least likely for the same reason it was fake - "too perfect". Amanda thought that Graphs 1 and 4 were most likely and Graphs 2 and 3 least likely. There is evidence in our conversation around the most likely/least likely phrasing that Amanda also attended to the ends of the distribution. Amanda focused on the number of handfuls containing nine and ten red candies, and the number of handfuls containing two, three or four red candies. Amanda seemed to expect fewer handfuls containing nine and ten red candies than she expected containing two, three, or four red candies.

Interview 1: Real/Fake Task:

Interviewer: Okay, so why are Graphs 1 and 4 more likely to happen for you than the other two graphs?
Amanda: I think for the same reasons I felt like two and three were the fakes. Graph 2 doesn't have a lot of variation occurring....We are incredibly heavy lumped in the 9 and 10 reds. And I'm uncomfortable with that, even though I realize in Graph 4 we have quite a concentration on the 9 red candies. But in Graph 2, nothing is going on below 4, which makes me uncomfortable.
Interviewer: Why?
Amanda: Because I think it should, something... at least one observation below 4 should occur.

In this excerpt Amanda decides that Graph 2 is less likely to occur because of the large number of handfuls containing nine and ten red candies. Amanda is also surprised that there are no handfuls containing fewer than four red candies in Graph 2.

I wanted to know if Amanda was relating the reasoning she employed for the Prediction Task to the Real/Fake Task. In particular, I wanted to know if she was thinking back to her own predictions or to the binomial probability structure that she used as a means to justify her predictions. I questioned Amanda about how her predictions for the Prediction Task related to her identifications for the real/fake graphs in order to see in what way these tasks were connected for her.

Interview 1: Real/Fake Task:

> **Interviewer:** So if I go back to your predictions [*I bring back the prediction task where Amanda made some predictions for 50 samples of 10 with the same population of candies as the real/fake*] you have one observation at four and then you have a prediction of one for 2 red candies and a prediction of one for 3 red candies and nothing else. So when you place these one's here, you're pretty certain that at least one of them is going to happen? Because when I see this I could also think, 2 red candies you've only got a one here. It might happen, but it might not. It's a low probability.
> **Amanda:** Well yes, I'm expecting to see something down here. Especially taken in conjunction with the fact that I have, how many are piled here on 9? A lot. Twelve, oh, 11 in the 9 slots on Graph 2, and 6 in the 10 slot. And I feel like this is a little disproportionate. I've got nothing here [*in 2's, 3's and 4's*] and a lot going on at 9 and 10. And I would feel more comfortable. Watch this. This is just going to be awful. If I removed some off 9 and 10 and moved them over here [*to the 2, 3 and 4 red candy slots*] so that it looked more like Graph 3. The theoretical one [*laughs*] that I think is implausible…. I'm having a battle in my head about theoretically what I expect to happen, which would look like Graph 3, and reasonably in practice what I have seen happen. Okay, I have not performed this exact experiment, but I've spent many, many hours drawing samples on a computer and seeing what they look like. And I'm talking about way more than 50 samples and they're always a little quirky. So what's occurring

here is a battle between what I theoretically know should be going on and what I realistically know should be going on.

I believe that this excerpt contains some important insight into Amanda's thinking. First, in this exchange, Amanda indicated that she was having a difficult time resolving her understanding of the theoretical model with her understanding of experimental data (see last highlighted utterance in the previous excerpt). Second, there is some indication in her utterances that Amanda believed there should be more balance between the number of handfuls containing nine or ten red candies and the number of handfuls containing two, three, or four red candies. In fact, in trying to create the balance that she expected to see between the two, three, and four versus the nine and ten red candies slots for Graph 2, Amanda realized that she ended up creating Graph 3, her image of the ideal graph. I believe that these utterances provide some evidence that Graph 3 is Amanda's image of the theoretical model, not just in *shape,* but in *spread* too. I conjecture that Amanda is struggling with the balance between sample representativeness and sample variability, and that she may not have a sufficiently strong sense of bounded variability. Although Amanda recognized the underlying probability structure as binomial and used that structure to help her calculate the expected values, she made her predictions more spread out than the binomial probability function indicates as likely. Amanda's predictions ranged from one handful containing two red candies to three handfuls containing ten red candies. Further, in the previous excerpt, Amanda indicated that she *expects* to see at least one handful containing fewer than four red candies. Graph 3 was manufactured to have an

unusually large number of handfuls, six out of 50, containing four or fewer red

candies. Yet Amanda's utterance indicated that she was drawn to Graph 3 not only in

terms of its shape, but precisely because it contains some draws with four or fewer red

candies.

The internal struggle that Amanda expressed in her examination of the real/fake

graphs between the two, three, and four red candies range, versus the nine and ten red

candies range can also be seen in her response to the hypothetical student predictions

on the Prediction Task. In the following exchange, I asked Amanda about her opinions

regarding Hypothetical Student 1's prediction (see Table 17).

Table 17: Hypothetical Student 1's Predictions

Number of Red Candies in Handfuls of 10	Hypothetical Predictions Student 1
0	0
1	0
2	0
3	0
4	0
5	5
6	9
7	15
8	18
9	3
10	0
TOTAL	50

Interview 1: Prediction Task:

Interviewer: Can I ask what's more troubling to you, the prediction of 0 at the 10 red candies place or the prediction of 0's here at the 0 through 4 red candies. Or is it equal?
Amanda: More troubling to me is the 0's in the 3 to 4 red candies region. I think. Yeah, 0's in the 3 to 4 region are more troubling to me because conceptually this is how it works for me. I see this giant bowl and its all speckled red and yellow. It's just as difficult for me to think of reaching in

and drawing 10 red candies as it is for me reaching in and grabbing 0 red candies. So it's less troubling for me that they don't feel that that is a reasonable outcome so they mark it 0, that's not going to happen. But I'm perfectly comfortable thinking I'm going to reach into this bowl and grab 4 red candies and 6 yellow. And that they don't mark that at all, that they anticipate 0 of those outcomes.

In this excerpt Amanda indicated that she imagined getting a handful with 10 red candies is as unlikely as getting a handful with 0 red candies. From my perspective, it appeared that Amanda was ignoring the ratio of red candies to yellow candies in the jar. Amanda's reasoning appeared inconsistent because she showed knowledge of the ratio of red to yellow candies and understood the underlying binomial probability structure, yet she appeared to expect to see as many or more handfuls containing two, three, or four red candies as handfuls containing nine or ten red candies. However, it seems that because Amanda's own concept image of the theoretical model was more spread out than the actual theoretical model, her reasoning was consistent from her perspective.

I continued to question her on how she visualized each end of the distribution, and asked her to make reference back to her own predictions. After considering these questions, her thinking changed slightly. Amanda indicated that it would be more difficult to get a handful with zero red candies than a handful with ten red candies. However, I think that Amanda still struggled with her image for pulling out two, three, or four red candies versus nine or ten red candies.

Interview 1: Prediction Task:

Interviewer: So from your perspective, how you imagine the situation, do you have a harder time imagining reaching in and pulling out 1 red candy or 10 red candies?

Amanda: Yes, I have a difficult time imagining either one of those scenarios happening, but I have a harder time with 1 red versus 10.

Interviewer: Because of the likelihoods?

Amanda: Right, when I imagine my big bowl of candies and how red it's going to look.

Interviewer: And what about 3 red candies in your handful or 10 red candies?

Amanda: I have a harder time imagining 10 red candies.

Interviewer: Okay. So you're thinking it's more likely I'm going to get 3 red candies?

Amanda: Umm, just in a visualizing sense. Yeah.

Interviewer: Okay, so when you say that, in a visualizing sense, that might not have anything to do with how the actual theoretical probabilities work out?

Amanda: Right, exactly.

Interviewer: Okay. Do you feel like you have a tension between the actual probabilities and this visualizing sense?

Amanda: Certainly. Certainly. And actually when you phrase it in terms of do you feel like it would be more likely, then instantly my gut reaction is to say, 'well I can't say that because I would have to sit down and calculate probabilities'. Umm, but just in my mind's eye visualizing way.

Interviewer: Okay. And what about 4 red candies versus 10 red candies? Or 4 red candies versus 9 red candies?

Amanda: I have a harder time imagining 10 versus 4. Umm, and I don't know about the 4 versus 9. I don't know why, but I don't know that one of those is harder for me to imagine. They might be more on par with each other for some reason.

This exchange shows that Amanda is thinking more about the proportion of reds to yellows in the jar because she made explicit reference to how red the jar looks. For this reason, she indicates that she more easily visualizes getting a handful containing all ten reds than she can see getting a handful with zero or one red candy. Yet, the strength of that image did not aid her in visualizing pulling out a handful of three or four red candies versus ten red candies. In fact, the probability for pulling out ten red

candies is larger than for three or four red candies, and ten red candies is closer to the expected value than three or four red candies. Yet, Amanda's image of the situation makes ten an extreme, and thus harder for her to visualize than three or four red candies.

Amanda also recognized that her image of the situation might, in fact, be different than how the actual computed probabilities would work out. The previous exchange sheds light into Amanda's image of the candy jar context and how that image influenced her own prediction and her identifications for the real/fake graphs. I believe the difference between Amanda's image of the theoretical model and that of the actual theoretical model is in part what led to the inability to resolve the tension she experienced between the theoretical model and the experimental data. This is why she rejected Graph 2 as plausible because she visualizes more handfuls containing three and four red candies than containing nine and ten red candies. That is, she envisions more variability than is likely to occur in a sampling situation. Although the ideal graph for this population would yield a shape similar to Graph 3, it would not have as high a concentration for two, three, and four red candies, yet this distinction was not part of Amanda's concept image (in the sense of Tall & Vinner, 1981).

Figure 21 provides a map of my interpretation of Amanda's reasoning path based on applying the conceptual framework and the continuum scale, developed in the conceptual analysis section, with Amanda's own image of the theoretical model in the candy jar context. The excerpts in this section provide strong evidence that Amanda focused on the attributes of shape and spread in her examination of the real/fake

graphs. Yet, she did not appear to apply these attributes together to provide an informal distributional argument. Rather, she appeared to focus on one attribute at a time. The coordination of multiple attributes of a distribution was difficult for her in this task. Rubin et al. (1991) observed this same difficulty in high school students. Amanda primarily used shape in her discussion for why she believed Graph 3 was made-up (or unlikely) and Graphs 1 and 4 were real (see conclusions in her reasoning path in Figure 21). Amanda primarily used spread in her discussion for why she believed Graph 2 was made-up (see conclusions in her reasoning path in Figure 21). There was no contradiction for Amanda because her image of shape matched up with the theoretical, but her image of spread did not. In addition, she expected more deviation from the theoretical model in shape, but less deviation from the theoretical in spread.

Figure 21: Real/Fake Task – Amanda's Approach

Amanda's Reasoning Path:

Amanda Focused on Single Attributes of the Distribution

Shape

Spread

Amanda's image matches the theoretical model

AND

Amanda expects some difference in shape of experimental graphs from theoretical model

Amanda's image is wider than the theoretical model

AND

Amanda expects spread of experimental graphs to be more similar to theoretical

Reject Graph 3 – 'too perfect'
Accept Graphs 1 & 4 – variable frequencies
Undecided – Graph 2

Reject Graph 2 – Not enough handfuls with 4 or fewer red
Accept Graphs 1 & 4 – more handfuls containing 4 or fewer red and fewer handfuls with 9 & 10 red
Undecided – Graph 3

Amanda appeared to experience internal conflict between her image of the theoretical model and her image of the experimental data. As the reasoning path in Figure 21 shows, her shape argument alone would not suffice to make a decision for Graph 2, since the graph does not have severe changes in frequency and at the same

time the graph is not completely smooth. As a result, Amanda needed to consider another attribute of the distribution. By focusing on the spread and looking at the ends of the distribution, she was able to make a conclusion about Graph 2 that did not lead to any contradictions with her conclusions about Graphs 1 and 4. That is, because Graphs 1 and 4 have fewer handfuls containing nine's and ten's and more handfuls containing three's and four's they still appear plausible. The internal conflict arises when she considered her expectation for getting more three's and four's and fewer nine's and ten's in light of Graph 3. Amanda realized that in order for her to believe that Graph 2 is plausible, she would want to see some of the nine's and ten's shift to the three's and four's. At the same time, she recognized that such a shift would make Graph 2 look more like Graph 3, which she believed to be fabricated because of its perfect shape. As a result of Amanda's image of the theoretical as having more spread than it actually does, and her belief that the spread for the experimental data should match more closely to the theoretical, she experienced conflict in her examinations of Graphs 2 and 3.

Andy

When I asked Andy to rate the graphs from most likely to least likely he responded that Graph 3 was least likely for the same reason it was fake, "too perfect". It did not appear that Andy's reasoning changed as a result of the rephrasing.

Interview 1: Real/Fake Task:

> **Interviewer:** What if I asked this question a different way. Suppose I said these are 4 simulations that took place. Same scenario, 50 samples of size

10, 250 yellow, 750 red and I asked you to rank these from most likely to occur to least likely to occur. So if you had to rate these, what order would you put them in?

Andy: Okay, well again it's based on how I am translating this – what would I expect the simulation to look like?

Interviewer: Yeah.

Andy: Okay, so I'm doing a simulation and my uncertainty is, I don't know if I've coded it correctly. That may be a reasonable way to look at this….If I get something like this [*points to Graph 3*], I go no way. I did something wrong. Or if it's someone else's code, I'm wondering did you really do the simulation or are you just pulling the ideal case?

Andy still believed that Graph 3 was the least likely graph to happen in this experiment because he viewed it as the ideal, at least with respect to its shape. He also believed that Graph 4 was less likely to happen than Graphs 1 and 2, but he did not appear to feel as strongly about this choice. The next exchange shows that from Andy's perspective, Graph 4 seemed ideal except the gap at five red candies, which made it suspicious from his point of view.

Interview 1: Real/Fake Task:

Andy:…What's wrong with this one [*Graph 4*]? Well it's missing the 5. It's not too unreasonable. I don't know. It just still feels too perfect. Even though it's missing the 5, it feels like it's ideal except for the gap. Gosh, what happened to the 5? I was sort of expecting there to be a 5 and then there wasn't. So that's the only thing that feels genuine. The rest of it just doesn't feel genuine.

Andy suggested that Graphs 1 and 2 would be more likely. Yet when Andy says that Graph 4 "just doesn't feel genuine", there is strong evidence that Andy is using his beliefs about the experimental situation rather than his statistical knowledge to make his determination. The next exchange reveals that he could not decide which of

those two graphs would be the most likely, but that they each contained enough

anomalies from the theoretical model not to raise his suspicions.

Interview 1: Real/Fake Task:

Andy: The most expected simulation given the sizes of everything. Well let's see. I don't know between these two [*points to Graphs 1 and 2*]. I can't really tell between the two what would be the most likely. This one [*Graph 1*] seems, well the three is kind of small. It's far from the mean. This one [*Graph 2*] is sort of biased in funny way [*pointing to the dip at 8 red candies*]. I don't know. I'd have a difficult time telling between these two.
Interviewer: So you would rate these two as the most likely [*Graphs 1 and 2*]?
Andy: Yeah, those are the most likely. I would definitely think…. Do I suspect that, you know, if I see something like this or this [*Graphs 1 and 2*] that I've written the code wrong or something, and it's going to generate garbage. Would I be able to tell? Well if it generated all the same value, it's screwed up. I detect something is screwed up. But, if I get a picture like this [*points to Graph 2*], it's not ideal, but I never expected the ideal. I might not be too surprised, and think yeah, I coded it correctly. It's not triggering an, 'I better go and find the bug'.

This exchange suggests to me that Andy does not expect an exact match with the

theoretical model and he appears to be situated a bit more toward the inexact end of

the continuum. Further, Andy's image of the shape of the theoretical appears to map to

the shape of the actual theoretical model. However, from our conversations there was

not enough evidence to conclude if Andy's image of the other attributes of the

distribution, such as center or spread, matched the actual theoretical model. Figure 22

provides a mapping of my interpretation of Andy's reasoning path based on applying

the conceptual framework and the continuum scale, developed in the conceptual

analysis section, with Andy's own image of the theoretical model in the candy jar

context.

Figure 22: Real/Fake Task– Andy's Approach

Andy's Reasoning Path:

Andy focused on a single attribute of the distribution

Shape

Andy's image of the shape of the theoretical model matches the actual theoretical model

AND

Andy expects the shape of experimental data to vary more considerably compared to the shape of the theoretical model

Andy's conclusions:
- Graph 3 and 4 are less likely to occur in a simulation
- Graphs 1 and 2 are more likely to

Sandy

As previously discussed, Sandy initially focused on the shape of the graphs. On the survey, Sandy identified Graphs 1, 2, and 4 as 'real' because they had variability in their frequencies, but she identified Graph 3 as 'fake' because it was "too perfect" to be 'real'. As with Amanda, there is no evidence in Sandy's survey response or her initial responses in the interview that her decisions about the real/fake graphs were based on her knowledge of the underlying probability structure, which she used to answer the Prediction Task. After Sandy provided her initial response and interpretation of the Real/Fake Task, I directed our conversation back to the Prediction Task in order to understand whether or not this information played any role in her decision making process.

Interview 1: Real/Fake Task:

> **Interviewer:** Now before you talked about Student 1's predictions as unreasonable because they had nothing for 10 reds [***Referring back to the Prediction task***]. But this graph [***Graph 1***] has nothing for 10 reds, how come you didn't say this one was fake?
> **Sandy:** Again, because I'm thinking that experimentally that that might happen. I don't know because here I have only three [***referring back to her prediction for the number of handfuls with 10 red candies***].
> **Interviewer:** And what about all the ones down here [***pointing to the number of handfuls that have four or fewer reds in Graph 1***]?
> **Sandy:** Yeah, I don't know. I had a really hard time with these graphs.

Figure 23 provides a model for my interpretation of Sandy's reasoning at the start of our discussion of the Real/Fake Task.

Figure 23: Real/Fake Task – Sandy's 1ˢᵗ Approach

Sandy's reasoning path at the beginning of the interview conversation:

Sandy began with a focus on a single attribute of the distribution

Shape

Sandy's image of the shape of the theoretical matches the actual theoretical

AND

Sandy does not expect an exact match between the shape of experimental data and theoretical model

Conclusion: Graph 3 is made-up, graphs 1, 2, & 4 could be real
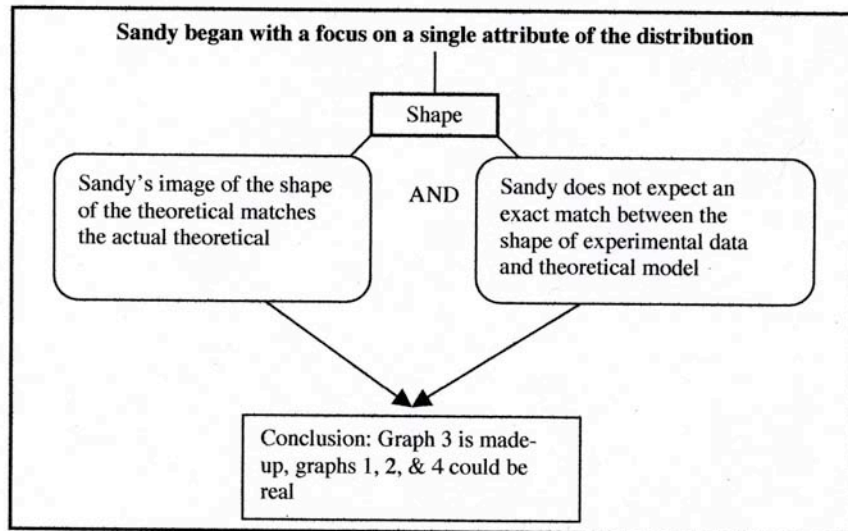
Like Amanda, Sandy also appeared to experience difficulty resolving her understanding of the theoretical model with the experimental graphs. However, when I rephrased the question in terms of rating the graphs from most likely to least likely,

Sandy felt more comfortable applying her knowledge of the underlying probability structure and comparing her own predictions based on the hypergeometric distribution to the real/fake graphs. Perhaps the most likely/least likely phrasing helped her to make a connection back to a probability statement.

Interview 1: Real/Fake Task:

> **Sandy:** Rate one through four? So I believe I would say Graph 2 is one [*most likely*]. [*Then Sandy marks Graph 4 as 2ⁿᵈ most likely followed by graph 1 and then last Graph 3*]…. So here [*Graph 2*] I can see experimentally that it's close to what I have [*comparing her predictions to Graph 2's outcomes*]. Again, Graph 3 is too perfect to be true. What I expect to see experimentally is something close to what I have, but not exactly like what I have. This means 0's here [*points to lower end of the graph*], peaking and then going down. Graph 1 is a three because this is three [*number of times 3 red candies occurred*], this is four [*number of times 4 red candies occurred*], pretty unlikely.

In this exchange, Sandy literally examined her predictions and compared those predictions with the graphs in the Real/Fake Task. By using her predictions, based on the theoretical probability structure, she was able to make conclusions about which types of graphs would be more likely than others. Her final choices agreed with the task design – Graphs 1 and 3 are made-up, and Graphs 2 and 4 are created by computer simulation. She also articulated that in general she expected outcomes *similar* to her own predictions, but *not exact* replicas. Figure 24 shows a model for my interpretation of Sandy's reasoning at the end of our discussion of the Real/Fake Task.

Figure 24: Real/Fake Task – Sandy's 2nd Approach

Sandy's reasoning path at the end of the interview conversation:

Sandy focused on the underlying probability structure

Formal hypergeometric probability distribution & her predictions for 50 trials of size 10 based on that model

Sandy's image of the theoretical model matches the actual theoretical model

AND

Sandy expects similar outputs between the experimental data and the theoretical model, but not exact matches

Conclusions: Graphs like 1 & 3 are less likely than graphs like 2 & 4

Joe

Joe's responses on the Real/Fake Task, as noted earlier, share some similarities to the responses of Amanda, Sandy, and Andy in that he was overly focused on the shape of Graph 3 and considered it the shape of the theoretical model. Yet, as the interview conversation around the Real/Fake Task continued, it becomes more evident that Joe's reasoning in this context also differs in some significant ways. Specifically, when I rephrased the real/fake wording to the most likely/least likely wording, my conversation with Joe went in a most unexpected direction. Rather than rating Graph 3 as least likely, as Amanda, Sandy, and Andy had done, Joe rated Graph 3 as most likely.

Interview 1: Real/Fake Task:

Joe: Oh yeah, that's a significantly different question. From most likely to least likely, I would do 1st most likely [*Graph 3*]. This is 2nd most likely [*Graph 2*]. This is 3rd most likely [*Graph 4*], and this is 4th most likely [*Graph 1*].

I was quite puzzled by Joe's ordering. His utterances in the previous exchange suggest that the graphs he considers made-up were the graphs he thought would be most likely to happen in the experiment. Joe's ordering of the graphs seemed to contradict his decisions about which graphs were 'real' and which were 'fake'. Yet, from Joe's perspective there was no contradiction because, as the next exchange reveals, he was thinking about the graphs "from the psychology of the student".

Interview 1: Real/Fake Task:

Joe: …If I'm seeing student work though, there's an additional level of context there. So seeing something like this [*Graph 3*] would make me suspicious. And this was posed in that context, so that's why there was the discrepancy of, I would be suspicious of this given it's student work versus I think this is the most likely given a computer generated model. Yeah.

It appears that Joe expected a student to know what the shape of the theoretical model would look like and that it would be the most likely shape to get in an experimental situation. Also worth mentioning is that in Joe's last utterance he seems to suggest that computer generated models will match theoretical distributions. This may indicate that Joe falls at the exact end of the continuum, expecting a match between experimental data and theoretical models.

As Joe continued to discuss the reasons for his ordering of the graphs, there also appeared to be a contradiction in his belief that there would not be many handfuls containing few reds and his belief that Graph 3 was most likely.

156

Interview 1: Real/Fake Task:

> **Interviewer:** In other words this is the least likely [***pointing to Graph 1***]?
> **Joe:** Yeah, just because there's 3 here [***3 handfuls with 3 red candies***] and 0 here [***0 handfuls with 10 red candies***].
> **Interviewer:** So here you're concentrating on this particular graph as least likely because you know from your intuition that it's less likely you're going to get a handful with 4 or 3 or 2 red candies and more likely you'll get something with 10, is that what you are saying?
> **Joe:** Yeah, yeah. If we assume the computer is a perfect random generator I would expect the most likely output to match this distribution that I have in my head.

Two points are worth noting here. The first being that Joe again provided explicit evidence that he expects experimental data to match up closely with the theoretical model. This can be seen in his last highlighted utterance where he suggests that if the "computer is a perfect random generator" then it will produce simulations that match the distribution in his head. The second interesting component in this exchange is that Joe rated Graph 1 as least likely because it had too many handfuls with two, three or four red candies and not enough with all 10 red candies. Yet, he did not use this same reasoning to suggest Graph 3 is less likely. Instead Joe argues that Graph 3 is most likely because it matches the shape of the theoretical distribution he has in his mind's eye. Further questioning provided possible reasons why this apparent contradiction was not a contradiction in Joe's mind.

Interview 1: Real/Fake Task:

> **Interviewer:** So let me ask you this. When you made your decision on Graph 1 you focused on the number of handfuls with 3 red candies versus the number of handfuls with 10 red candies. But if I look at Graph 3, there's a pretty similar number that are down here [***I point to Graph 3 at the 2, 3 and 4 red candies spot***] that are down here [***pointing back to Graph 1 in the 3 and 4 red candies spot***]. So how come Graph 3 makes it

to your most likely category, whereas Graph 1 made it to your least likely and one of your reasons was the number that were at 4 or fewer reds in a handful? Does that make sense what I'm saying?

Joe: Yeah, that does make sense. So, if we look here [*Graph 3*] at 4's and below there are 7 of them. And here [*Graph 1*] at 4's and below there are 6 of them. And those are very close numbers, just one away [*laughs*]…. But here [*Graph 1*] we have kind of a divot [*at the 5 red candies spot*] and then it goes up [*at the 4 red candies spot*] then it drops to nothing [*as Joe says this he draws curve over Graph 1*]. And it's not unlikely, I just think this [*Graph 3*] is more likely [*draws nice curve over Graph 3*].

Interviewer: So it's more about the shape for you?

Joe: It's more about the shape for me because of my minimal experience with these. So the tool I have is what kind of shape do I expect.

I believe that this excerpt reveals some interesting insight into Joe's thinking and the tensions that he experienced between the theoretical model and the experimental data. Although Joe recognized the ratio of red to yellow candies and had strong intuitions about the chances of pulling out 10 reds versus 3 or 4 reds, his predominant form of reasoning was based on shape. In addition, Joe's attention to shape was with respect to changes in frequency, similar to that of Amanda, Sandy, and Andy. Joe attended to the 'ups and downs' in the frequency in Graph 1 from the three, to the four, to the five red candies spot. In contrast to Amanda, Sandy, and Andy, Joe believed that 'ups and downs' in frequency was an indication of unlikely graphs and that smooth curves, such as that exhibited in Graph 3, was an indication of likely graphs. That is to say, it appears that Joe believed that the theoretical model was the one that was most likely to happen in the experiment and he primarily focused on the shape of that theoretical model. Figure 25 shows a model for my interpretation of Joe's reasoning at the end of our discussion of the Real/Fake Task.

Figure 25: Real/Fake Task – Joe's Approach

Joe's Reasoning Path:

Joe Focuses on the Two Attributes of the Distribution

Shape

Spread

Joe's image of the shape of the theoretical model matches actual theoretical model

AND

Joe expects experimental data to produce *exact* match in shape to theoretical

Joe's image of the spread for the theoretical model matches with the actual theoretical model

AND

Joe expects experimental data to have a similar spread as the theoretical model, but allows for more variation here than in shape

Joe's Conclusion:
- Graph 3 is most likely – shape fits with the theoretical
- Graph 1 is least likely – spread deviates too much from the theoretical
- Graph 2 is more likely since it's shape is closest to graph 3, & range is closer to theoretical, from Joe's perspective
- Graph 4 is less likely since it's shape deviates more from graph 3, & has a range that deviates more from theoretical, from Joe's perspective

4.1.4 Summary for Prediction Task and Real/Fake Task

The difference in the way TAs in this study appeared to reason about the Prediction Task as compared to the Real/Fake Task is a particularly compelling result of my study. As highlighted throughout this analysis, most of these TAs appeared to be able to reason using formal and/or informal attributes of the distribution in order to justify their predictions for the Prediction Task. However, few TAs used such

159

arguments to justify their decisions in the Real/Fake Task. TAs tended to use shape as their primary justification for why certain graphs were 'fake' or less likely to occur and why certain graphs were 'real' or more likely to occur. Some TAs also employed a range/spread argument when they could not make a decision based solely on shape. Yet, these alternative arguments were only secondary, and if such an argument contradicted with their image for the shape of the experimental data, shape appeared to hold more power in their final decision. For example, when Joe decided Graph 1 was unlikely because of the large number of handfuls with four or fewer reds, he abandoned that reasoning path and pursued a purely shape argument to conclude that Graph 3 was likely.

The salience of shape over the other attributes of distribution seem, in part, related to how these TAs thought about variability in this context. Rather than focusing on the concept of bounded variability, TAs appeared to expect graphs that have more unevenness in frequencies to be more likely. The Real/Fake Task was not a routine application of a binomial or hypergeometric model for these TAs, which may explain why they had a difficult time quantifying their expectations for the experimental sampling distributions. As a result, TAs tended to rely on their subjective beliefs for what might happen in the experiment. I suspect that if TAs had practice actually engaging in such an experiment, their subjective beliefs would probably tend toward a strong notion of bounded variability in a more statistical sense. Finally, the primary focus on shape in justifying real/fake identifications could be the result of the task

providing graphical displays of data. It would be interesting to see how TAs would respond to this same task if the data were presented in tabular form.

Also compelling is that TAs' decision-making process for identifying real/fake graphs appeared to be a function of their images of the theoretical model compared to the actual model, and how closely they expected experimental data to map to the underlying probability distribution model. It appeared that of the TAs I interviewed, most of their images of the shape of the theoretical model matched closely to the shape of the actual theoretical model. Unfortunately, for the most part it remained unclear how their images of the other attributes of the distribution mapped to the actual theoretical distribution. Finally, there was some variation from TA to TA in their individual expectations for how close a mapping they expected between the experimental data and the theoretical model. For instance, on the one hand, Amanda, Sandy, and Andy seemed to expect something similar, but not exact, particularly with respect to shape. On the other hand, Joe expected a close match in shape between the experimental data and the theoretical model.

A final point worth mentioning is that TAs expressed some distress in making their decisions about the real/fake graphs. In fact some TAs, like Amanda, explicitly expressed their decision-making process as a mental battle between what they expected from the theoretical model and what they expected or experienced in experimental data. This tension appeared to remain as TAs continued to interact with the interview tasks involving experimental data, like the Mystery Mixture Task, for instance.

4.2 Mystery Mixture Task

The Mystery Mixture Task followed the Prediction and Real/Fake Task in the first interview. The scenario was similar to that of the Prediction and Real/Fake Task: there were 1000 red and yellow candies in a jar and four groups of students took 50 samples of size 10. For each sample of size 10 the groups recorded the number of reds, replaced the candies in the jar and remixed before drawing the next sample of size 10. TAs were shown the four student recorded graphs (see Figure 26), one from each group, and were asked to use that information to predict the number of red candies in the jar. That is, I asked TAs to make an inference about the value of a population parameter on the basis of empirical sampling distributions. The jar contained 350 red and 650 yellow candies, but this information was not disclosed to TAs.

Figure 26: Mystery Mixture Graphs

162
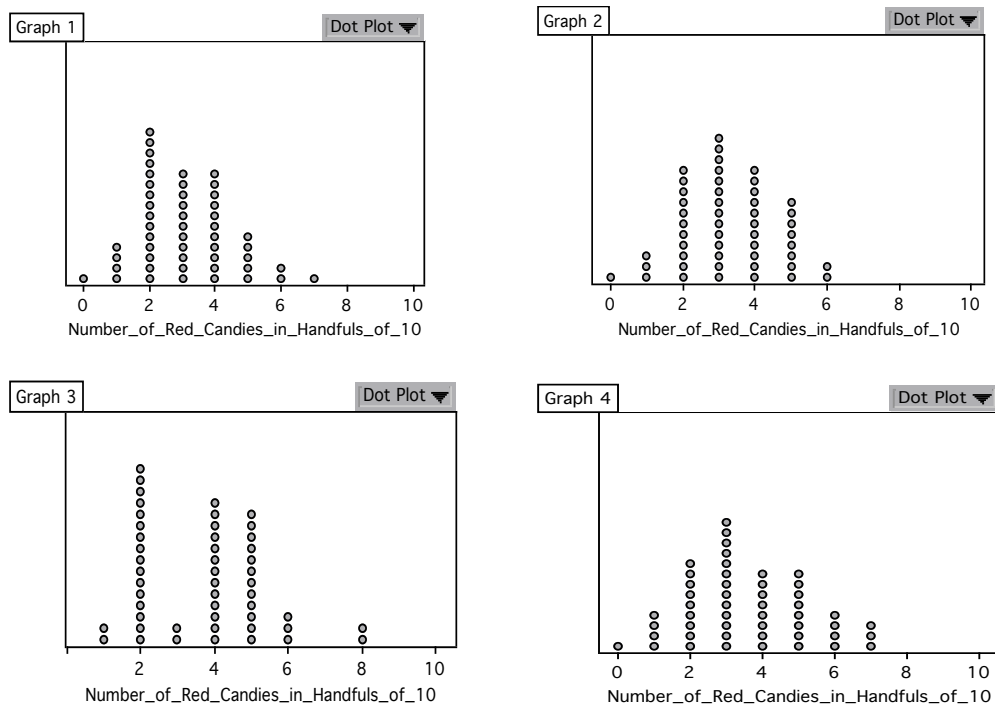
The purpose of this task was to provide TAs another opportunity to interact with experimental sampling distributions and thus afford myself another opportunity to investigate the ways in which TAs reason about the graphs of experimental data.

4.2.1 Conceptual Analysis of the Mystery Mixture Task

In the Mystery Mixture Task, the underlying population proportion is unknown so one cannot routinely apply the binomial or hypergeometric probability distributions models and arrive at an exact solution. The TAs in this study focused on one or more attributes of the distribution to determine an estimator for the population proportion. As with the Prediction and Real/Fake Tasks, a TA may focus on a single attribute of the distribution or multiple attributes. Further, a TA may employ a single graph or multiple graphs to justify their point estimates. Figure 27 shows my conceptual framework for how TAs reasoned on the Mystery Mixture Task.

Figure 27: Conceptual Framework for Mystery Mixture Task

TAs' Reasoning about the Mystery Mixture Task

```
┌─────────────────────────────────────────────────────────────┐
│   TAs' Estimate is Based on Information from a Single Graph   │
│              /                              \                 │
│          Shape                           Center              │
│                                             |                │
│                                           Mode               │
│                                             |                │
│                                      Median or Mean          │
└─────────────────────────────────────────────────────────────┘
                              |
┌─────────────────────────────────────────────────────────────┐
│   TAs' Estimate Incorporates Information from Multiple Graphs │
│  (Implicit knowledge that a better estimate can be obtained   │
│                      with more information)                   │
│                                                               │
│                           Center                              │
│                             |                                 │
│                       Modal Average                           │
│                             |                                 │
│                    Median or Mean Average                     │
│                             |                                 │
│      Focus on Multiple Measures of Center & Average of Averages │
└─────────────────────────────────────────────────────────────┘
```

Notice that the attributes of shape and center are the only attributes displayed in this conceptual framework. That is because none of the TAs I interviewed focused on the spread of the distribution. I suspect that the reason for this is the wording of the task. In this task, TAs were asked to provide a population parameter for the number of red candies in the jar, leading to a natural focus on measures of center.

Table 18 shows the mode, median, and mean for each graph, as well as the average of the three centers over the four graphs. Of course, the mode is the quickest method for visually assessing the population parameter, but within the statistical community the mode is generally considered a less sophisticated and less accurate predictor for the population parameter. The median and the mean for each graph provide more accurate estimations of the population parameter (μ=3.5).

Table 18: Measures of Center for Mystery Mixture Graphs

	Mode	Median	Mean
Graph 1	2	3	3.1
Graph 2	3	3	3.26
Graph 3	2	4	3.72
Graph 4	3	3	3.56
Average	2.5	3.25	3.41

The median, being the middle data value, would not take too long to calculate for this task – one could simply count in 25 data points from the left or the right. The mean would take more time to compute for this task, but could be visually estimated by assessing where the balancing point would be for the data. However, the graphs obtained from each of the 50 trials of 10 also provide an indication that the distribution is right-skewed. This observation indicates that mean, or center of mass, is pulled to the right because of the long tail of the distribution. Figure 28 shows a right-skew distribution and the location of the median, $\tilde{\mu}$, and the mean, μ.

Figure 28: Right-skewed distribution
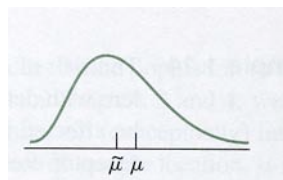
4.2.2 TA Thinking and Reasoning about the Mystery Mixture Task

My analysis of the responses of four of the TAs I interviewed provide more support for my conjecture that TAs' experience tension in working with experimental data, and as a result, tend to resort to less sophisticated types of reasoning, such as shape or mode. I begin the discussion with the initial reactions of Amanda and Joe. Amanda and Joe were mainly focused on the modes of each of the graphs and used a modal average to provide an estimate for the number of red candies in the jar. I follow with a discussion of Sandy's reasoning. Sandy used a shape argument and indicated that the shape was reverse from that of the Prediction and Real/Fake Task. Ultimately though, Sandy did not want to provide a specific answer to this problem because without a calculation it would not be precise. Finally, I end this section with a discussion of Andy's reasoning. Andy focused on the means of the graphs and appeared to approximate the average of the four means.

Amanda & Joe: Modal Average

The following exchange shows Amanda's initial reaction to the Mystery Mixture Task. It appears that she is inclined toward computing a modal average. Since two of the graphs have a mode of three and two graphs have a mode of two, Amanda provided an estimate of 250 red candies.

Interview 2: Mystery Mixture Task:

> **Amanda:** These are awfully conflicting graphs.
> **Interviewer:** Why?
> **Amanda:** 'Cause we have two where three is clearly dominant, and two where two is clearly dominant.

166

Interviewer: So you're looking at the modes of the graphs?
Amanda: Right. Sorry the modes. So in Graphs 2 and 4 we've got a mode of 3 and in Graphs 1 and 3 we've got a mode of 2. Alright, I'm going for 250 reds, 750 yellows.

As Amanda continued to engage with the task her attention wandered from the two and three red candies spot to four red candies. However, any information that she gleaned from observing the other outcomes in these four empirical sampling distributions did not appear to prove convincing enough for her to ultimately change her prediction.

Interview 2: Mystery Mixture Task:

Amanda: And I would even be inclined to, no see they're really conflicting for me because I would also be inclined to say 300 red, 700 yellow. Or maybe somewhere in between those two predictions. We clearly have concentration at 2 and 3. And oh kind of 4. Which is drawing me to the 300 versus 700. But then that doesn't explain these graphs [*Graphs 1 and 3*] that have modes at 2. Which is why I said 250.
Interviewer: So 250 is you sort of splitting the two modes?
Amanda: Right. Well none of these graphs has a mode at 4, which is pushing me to having between 250 and 300 or maybe even 200 to 300 red. [*Long pause*]. Graph 2 we have some symmetry around 3, but taking this in conjunction with the other graphs I'm still kind of leaning toward. Sorry this is all the thought that went into picking 250 versus 750. And Graph 3 I'm going to take as an anomaly that we only got 2 readings for 3 [*red candies*]. I think that's a freakishly bad, unfortunate sample and the other 3 graphs are leading me between 250 and 300 red.

In the previous exchange, Amanda did take note of the large concentration of handfuls containing four red candies, yet because the mode was never located at four, she decided to ignore that information in the end. Amanda also provided an informal confidence interval of 200 to 300, or 250 to 300 red candies based on the modal

values. She was moderately troubled by Graph 3 having so few handfuls with three red candies, but she considered the situation as an anomaly.

The following exchange reveals that Joe's initial prediction and justification is similar to Amanda's.

Interview 2: Mystery Mixture Task:

Joe: I look at the curves and there's a spike at 3, spike at 3, [***Joe points to the modes of Graphs 2 and 4***] spike at 2, and a spike at 2 [***Joe points to the modes of Graphs 1 and 3***]. So maybe there's more like 250 [***reds***].
Interviewer: So you're looking at the spikes and basically you're telling me that on Graph 2 and 4 the spikes occur at 3, those would be the modes? And then on Graph 1, the mode is 2?
Joe: And Graph 3 the mode's at 2.
Interviewer: And you are averaging those two?
Joe: Right. Yeah, and actually the first thing I looked at was the shape of these curves and how there's. It's kind of a peak right around in here [***draws curves over each of the graphs and appears to point to the 2's and 3's for the peak***].
Interviewer: 2's and 3's?
Joe: Yeah. Right around in here [***Joe points to the peaks for each of the four graphs***]. I'd say it's the opposite, 250 red to 750 yellow.

It appears that Joe used the words 'spike' and 'peak' to talk generally about the modes of each of the graphs and perhaps even where the largest clump of data fell. It appears that he averaged the modes of two and three to come up with his estimate of 250 red.

Thus, Amanda and Joe focused primarily on the modes of each graph and took the average mode for the four graphs to justify their predictions of 250 red candies in the jar. They reasoned using all four graphs, but only one attribute of the distribution – center. Also, they focused on only one measure of center, the mode. Again, the focus on measures of center is not surprising given that the task asks for a point estimate. However, the fact that Amanda and Joe focused primarily on the modes to make their

decisions is surprising given their statistics backgrounds. Perhaps the graphical

displays of data played a hand in their focus on modes since it is easy to detect in a

visual presentation of data. It would be interesting to provide a similar question to TAs

with data presented in tables to see if this presentation would change their approach to

the problem.

Sandy: Focused on Shape

From the beginning of this task, Sandy appeared visibly stressed. Sandy liked

performing exact calculations; she did not like making estimations. Sandy initially

began the conversation by discussing the shape of the graphs.

Interview 2: Mystery Mixture Task:

Interviewer: This is a Mystery Mixture Task. It doesn't tell you what the mix of candies are in the jar. We only know that there are 1000 candies, red and yellow. They're graphing the number of reds just how we graphed here [*referring to the Real/Fake Task*]…
Sandy: So now I have no clue. So now it seems that you get it skewed to the right. So before I had 750 red. Okay, I had 750 red. I got it skewed to the left and 250 yellow. I'm just thinking that to see it like that I should have red smaller than yellow.
…
Interviewer: So would it be the same mix, but the reverse? 750 yellow and 250 red?
Sandy: No, I'm not sure about that I cannot. I can see that for example here [*Sandy points to Graph 4*] it peaks at 2 and 3. If I reverse it, I knew that before it peaked at 7 and 8 [*in the Real/Fake Task*]. Just by looking at this graph [*Sandy points to Graph 4 again*]. So it might be, it might be the same if I'm looking just here you know [*Sandy points to Graph 4*].

Sandy started the conversation saying that she "had no clue". She also reiterated to me

that she did not have experience working with experimental data. It appears that

Sandy's attention was focused on the shape of the four graphs. She noticed the skewed

right distribution and that Graph 4 had peaks at two and three. Sandy compared the real/fake graphs to Graph 4 in the Mystery Mixture Task. She noted that the real/fake graphs peaked at seven and eight red candies and that there were more red candies in the jar than yellow. Sandy went on to suggest that since Graph 4 in the Mystery Mixture Task peaked at two and three red candies that perhaps this jar contained fewer yellow candies than red. I thought that Sandy might be inclined to reverse the mixtures in the jar based on her comparisons between the two tasks. Yet, Sandy was careful not to estimate the number of red in the jar. She indicated that the mixture could be the reverse of the real/fake mixture, but she could not say for sure.

At this point in the conversation Sandy was only using one graph, Graph 4, to reason about the mixture in the jar. I asked Sandy about the other graphs and how she might come to a decision based on all the information provided. I believe the following exchange provides an indication that Sandy was not comfortable gleaning information about the mixture using all four graphs provided.

Interview 2: Mystery Mixture Task:

> **Interviewer:** What if you look at all of them together [*referring to all 4 graphs on the page*]?
> **Sandy:** This one [*pointing to Graph 3 at the dip at 3 red candies*] doesn't tell me anything. And here this one also [*points at Graph 1*] because it's like 4 is close to 3 [*referring to Graph 1, the number of samples with 3 red candies is equivalent to the number of samples with four red candies*]. You see. So I'm not sure.
> **Interviewer:** What if instead of trying to predict the exact amount you were just trying to come up with a confidence interval, like the number of red is between such and such, that you would feel pretty confident you would capture it.
> **Sandy:** I cannot do that.
> **Interviewer:** No?

Sandy: No and I don't like to do that.
Interviewer: Because it's too iffy?
Sandy: Exactly. I don't like. I, so how I approach problems is give me the information I will apply the formulas, I will give you the answer. And I'm confident of doing that.

From this exchange it appears that Sandy disregarded Graph 3 as providing useful information because of the unusually low number of outcomes at 3 red candies in comparison to the other graphs. Sandy does not specifically state this as her reason, but she pointed to the dip at three red candies when she indicated that Graph 3 did not provide her any information. Other TAs mentioned the gap at three red candies on Graph 3 as troubling, but they did not seem as quick to discredit it. Sandy also suggested that Graph 1 did not provide her any useful information because the number of outcomes for three red candies was the same as for four red candies. It is unfortunate that I did not probe more into this comment because it is unclear why Sandy viewed equal outcomes of three and four red candies as troubling, especially in light of the fact that Graph 2 has an equal number of outcomes at two and four red candies and Graph 3 has an equal number of outcomes at four and five red candies. It appears that the four experimental graphs provided conflicting information that paralyzed Sandy's ability to obtain useful information from the data. Sandy was not able to coordinate her actions on one graph to all four graphs. The last utterances suggest that Sandy is not interested in making estimations. She prefers to apply a specific formula and obtain an answer.

Sandy did not appear interested in working with the Mystery Mixture Task as it was, so she tried to turn the problem into something she felt more comfortable with.

Interview 2: Mystery Mixture Task:

> **Sandy:** What you can think is like you have different populations. Four populations like that [***pointing to each of the graphs***]. Four different populations and you go and in each population you collect a sample and you want to see if the average of people who have attribute one in the first population is equal to the average of people who have attribute one in the second population, third and fourth. You can do that, that type of test.

The previous exchange provides evidence that Sandy appeared to be comfortable with thinking that each of the four graphs represented samples from different populations and the problem of interest was to test whether the means for each population are the same. I found it interesting that Sandy did not want to think that all four graphs were from the same population, so I continued to question her on this point. The next exchange provides some insight into why having four collections from the same population troubled Sandy.

> Interview 2: Mystery Mixture Task:

> **Interviewer**: What if these are all from the same population?
> **Sandy:** All from the same population and you collect?
> **Interviewer:** Yeah, like the information in Graph 1 is what I collected and the information in Graph 2 is what you collected. It's all about the same group of people. With all of that information could we get an idea about what percentage of the population had that attribute? Like what percentage of the population was red versus…
> **Sandy:** I'm confused right now because normally if you want to conduct a test you have one population. You'll go ahead and collect one sample.
> **Interviewer:** But in stats don't we talk about repeated sampling?
> **Sandy:** No…. If you want to conduct this test, you are looking for the proportion of people who have red hair, you know. You want to know if the proportion of people who have red hair is equal to say 75% or different than 0.75. How do you conduct this test? You go and collect one sample…. So you go and collect *one* sample, so right now if you are telling me you collect four samples I don't know what to do with four samples. [***Laughs***] You know.

Interviewer: Yeah. I understand that because we always do one sample right? But if we had more than one sample? Wouldn't that give us more information?
Sandy: No.

From Sandy's statistical experience, collecting multiple samples is not an appropriate statistical technique. During the exchange, Sandy suggested that she did not know what to do with four samples. That utterance also suggests that Sandy may have thought about each one of the graphs as one sample and not as 50 samples of size 10. Sandy did not believe more samples meant more information. It appears that Sandy does not have a strong conception of sampling distributions in the point estimate context. This lack of conception of sampling distributions is consistent with her reasoning in the Unusual Sample Task (see Chapter 5). Her experience of taking a single sample and performing a test on that sample appears to have influenced how she thinks about data collection and the information that can be gleaned from the data. For Sandy, if there are multiple samples the information could be conflicting and/or it will be unclear how to perform the appropriate statistical test.

Andy: Estimating the mean

Andy appeared to focus his attention on the attribute of center. But unlike Amanda and Joe, the mean, not the mode, was Andy's choice for estimating the number of red candies in the jar.

Interview 2: Mystery Mixture Task:

Andy: Okay. My first response if you want me to talk it through, since we're on video. The three, you know these [*Graphs, 1, 2 and 4*] are all suggesting if I were to draw the curve I'd say three-ish. Where's the mean?

I'm looking, that's about three [***Andy's mental estimate of the mean of Graph 1***]. Boy that sure looks like it's close to three [***mean of Graph 2***], man it's even symmetric about that [***symmetric about 3 red candies***]. Maybe a little more than 3 like 3.5 [***Andy's mental estimate of the mean for Graph 4***]. Maybe 350 reds [***points to Graph 4***] or 300 reds [***points to Graph 2***]. And then it's like boy this [***points to Graph 3***] really blows my theory. It's like what happened? Then again, it's like what on earth happened here? How do you get all these two's and four's, but the three's just sort of don't happen. Things happen. But maybe I've got to throw that out. It just seems so extreme and unlikely. If it happened by chance or maybe there was some other factor.

Interviewer: Okay, so you might throw Graph 3 out altogether?

Andy: Well I might. I don't know. I'm not in favor of throwing data out entirely if there's information there. I don't like throwing out information. So if there's a defect, one way to fill it in is to say just redistribute these and knock this one over a little bit. Because they're not saying it can't be three. This one [***Graph 3***] is consistent with the theory that it's 3 point, I don't know 3.5 or something. It's not inconsistent. It's just kind of funny. It's silent on whether it's 3 or not. Look at, what's the median. Even though this is a biased distribution and the median is not a great approximation for the mean, but it's not a bad place to start.

Andy did not appear to experience any tensions with the experimental graphs for this task. Unprompted, Andy was the only TA to discuss mean and median values for the graphs rather than the mode. Andy first provided an estimation of the means for each graph. His estimations resulted in him providing an approximation of 300 to 350 red candies (recall that the actual population parameter is 350 red candies). Andy also discussed the possibility of looking at the medians, especially for Graph 3. Andy did express concern about Graph 3. In particular, he was troubled by the low number of handfuls containing three red candies in light of the information provided by the other three graphs. Yet, this anomaly did not hinder Andy's ability to provide an estimation of the number of red candies in the jar. Andy suggested that this graph could be thrown out, or mentally redistributed by moving some of the outcomes at two and four

red candies to the three red candies spot. Also, Andy seemed to believe that Graph 3 was still consistent with the theory that the mean was between 3 and 3.5. This statement suggests that he may have been mentally estimating the balancing point for Graph 3, or thinking about where the median might fall for Graph 3.

Amanda: Pushed to Estimate Means

In the following exchange, I asked Amanda to describe any other elements that influenced her prediction for the Mystery Mixture.

Interview 2: Mystery Mixture Task:

Interviewer: Is there anything else that's influencing you? Any shape or spread ideas? And you talked a little bit about shape with the symmetry on Graph 2.
Amanda: A little bit of symmetry around the 3, right. They all have this similar distribution where the tail is on the right. Which you've had me staring at that other picture for so long, where it's the reverse of what my prediction was with the tail on the left. So that might be effecting what I'm thinking about, but really all I'm consciously processing right now are really where these concentrations are. I did momentarily make a conscious effort at examining the distributions across the other numbers. I can tell you right now that my brain froze up and I didn't know how to process that information so my brain went back to the concentration around the 2 and the 3. So I really did make an effort for a minute to consider 5, 6, and 7 [*red candies spots*] on these graphs, but it didn't feel. God I've never had to pick apart how I think about these things. But I instantly had a feeling of uncertainty about how to get any information about that, so I immediately went back to the mode.
Interviewer: So kind of on a gut level you could go back to the modes and be able to say something.
Amanda: Right, right.

The highlighted utterances suggest that Amanda did attend to the other outcomes in the distribution, but because she did not feel comfortable she returned to the modal average, where she felt confident using that information to make some sort of

prediction about the mixture of red and yellow candies. I think this excerpt provides

some indication of the tension Amanda experienced when she was trying to glean

information from experimental data. Amanda's comment "God I've never had to pick

apart how I think about these things" is also worth noting. It suggests that she has not

had to think deeply about "simple" statistical ideas, such as finding a point estimate

from some experimental data sets. Her comment is a strong indication that sampling

concepts and processes are, in fact, quite complex. Also, Amanda's comment suggests

that creating dialog that probes deeper into topics of sampling are important for

developing a more robust knowledge of sampling[22].

I asked Amanda what she would do if she could take the Mystery Mixture Task

home with her and was under no pressure to solve it. The highlighted utterances in the

following exchange indicate that Amanda would use more sophisticated forms of

reasoning about the graphs, such as identifying the means and medians, and/or

aggregating the data from all four graphs.

> Interview 2: Mystery Mixture Task:
>
> > **Interviewer:** Okay, if you had more time, or you were going to take this home with you, what kinds of things would you look at if you were going to ponder this for a little bit?
> > **Amanda:** I don't know. I'm trained to run diagnostics. So I might look at the average for each graph. Maybe the median for each graph and see how those compare to each other. Maybe the average of the averages. Because I think you know in some fashion you should be able to fuse all four graphs to be able to come up with a reasonable estimate of what's going on.

---

[22] This point will be discussed in more detail in Chapter 7 – directions for future research.

I pushed Amanda in the interview for a visual estimation of the means for the four graphs. As the following exchange reveals, she did not like to make estimations and preferred to provide exact answers.

Interview 2: Mystery Mixture Task:

Interviewer: If you had to visually, could you point to where you would see the means for each of the graphs? Or the medians, since you brought that up? Is that asking too much?
Amanda: Oh God. [***Long pause***]. You're talking to somebody who really has to, like it's important to me to sit down and calculate stuff. Like I'm really uncomfortable with saying, 'oh I think the mean is right about there'. [***Laughs***].

After pushing on Amanda a bit more to estimate the means for the four graphs she finally yields, but toward the end she was guessing more than estimating.

Interview 1: Mystery Mixture Task:

Amanda: Right now, I'm even freezing up on how the mean compares to the mode when you have skewness going on [***Amanda draws a sketch, see Figure 29***]. So the mean should be to the right of the mode when the tail is in the right. So, I don't do well under pressure. You're lucky you're not one of my professors or this would be dribble coming out of my mouth. And I also have a hard time with the visualization, knowing that it's stacked. I think the mean is, I'm going to slice it right here. So for reference that was Graph 1. I think maybe somewhere between 3.5 and 4. 3.5 and I'm sticking with it. Graph 2, 4. I don't know at this point I'm just trying to give you answers. Not uneducated answers, but I don't feel really confident about what I'm saying. 4.5 for Graph 3. Because I'm not counting x's or anything.

Figure 29: Amanda's sketch of skewed left distribution

Interviewer: But you are talking about these as being the means?

Amanda: Means. Yes. Not medians. I would really have to sit here and count *x*'s to do that. And I'm going to go with 4 again on Graph 4.

Interviewer: Okay. So doing that, does that change your initial prediction, or is this visualizing the means too fuzzy without calculations to make you change your predictions?

Amanda: Yeah, I don't really feel very comfortable, which is odd. I don't know which one of them would be more accurate in my estimation, but I feel really uncomfortable out of a bunch of *x*'s trying to select where the mean is. But if they all tended to focus around the 4 and if I felt with any confidence that that was correct [***long pause***]. No I'm going to leave it.

Table 19 shows a comparison for Amanda's estimations of the means for each of the four sampling distributions compared to the actual means of those distributions.

Table 19: Comparison of Amanda's estimation of means

Graphs	Amanda's estimation of the mean	Actual mean
1	3.5	3.1
2	4	3.26
3	4.5	3.72
4	4	3.56

Amanda's estimations for the means were not exact, and in particular in relation to Graph 3 her mean was quite high at 4.5 red candies. However, her estimations did indicate that the means were above three red candies. I thought that after Amanda estimated the means for each graph she would increase her estimate of the number of red candies in the jar. Yet, Amanda did not have enough confidence in the exactness of her estimates to change her prediction from 250 reds in the jar to something at least slightly larger, despite the fact that her mental estimates of the mean suggested that the population parameter was likely to be larger than 250 red candies. This suggests that Amanda had poor imagery of mean as a balance point (i.e., a center of mass). Amanda preferred to use the modes because she could calculate the modal average exactly.

4.2.3 Summary of the Mystery Mixture Task

There are two main points that deserve emphasis from my analysis of TAs'
thinking and reasoning about the Mystery Mixture Task. The first point is that three of
the four TAs experienced difficulties as they attempted to use the experimental data to
make a decision about the population of red and yellow candies. Like the Real/Fake
Task, the Mystery Mixture Task provided another context in which these TAs were
forced to confront experimental data and make decisions based on the experimental
information. On the one hand, Amanda, Joe, and Sandy all appeared to experience
some level of frustration and tension in trying to make sense of the four experimental
graphs. Amanda and Joe used a modal average and Sandy used a shape argument in
order to resolve the tension. By using mode as the measure of center, Amanda's and
Joe's predictions for the number of red candies differed from the population of red
candies by 100. Sandy never provided a definitive estimate. She would only suggest
that the skewed right shape implied more yellow candies than red. On the other hand,
Andy did not appear to experience the same frustration or tension that the other three
TAs experienced. He did not seem compelled to provide an exact mean value. Rather,
he made a visual approximation for each of the graphs at somewhere between 3 and
3.5, which allowed him to provide a strong overall estimate for the number of reds in
the jar.

Andy's experience with this task compared to the other three TAs leads to a
second point that deserves emphasis, estimation skills. I found that several TAs, both
on the survey and in the interviews, expressed distaste for estimation. This is

surprising as estimation is the crux of statistical inference. Yet, TAs are trained to use formal methods of estimation and the Mystery Mixture Task pushed them to make informal estimations. Several TAs on the survey mentioned that they used calculators on the Prediction Task in order to produce exact answers. Amanda did not feel confident in her estimates of the means for each graph in the Mystery Mixture Task. Also, at several different points during the three interviews, Sandy expressed a dislike for the interview tasks because she did not have an exact formula to follow (i.e., formulas that would provide 'exact' answers). I believe that this distaste for providing approximations is part of the graduate school culture. Mathematical and statistical coursework instills in TAs the desire to produce exact answers and to avoid estimation. Of course in many respects this is important in the work in which statisticians and mathematicians engage, but it is also important to have strong estimations skills. NCTM (2000) emphasizes the importance of computational fluency, strong estimation skills, and the sense of when it might be appropriate to provide estimates. The ability to mentally estimate the mean and/or median for each of the mystery mixture graphs allows one to provide a fairly accurate and quick approximation of the number of red candies in the jar. I believe that graduate mathematics and statistics courses are not likely to spend time working with experimental data, discussing processes for making decisions based on multiple samples or making use of estimation techniques. Thus, lack of experience with this type of problem is likely to account for the tensions these TAs experienced. I argue that TAs could benefit from professional development workshops in which they had

the opportunity to perform repeated sampling experiments, create graphs of experimental sampling distributions, and use the information gathered from a collection of samples to make inferences about the population from which the sample came.

4.3 Chapter 4 Conclusions

Taken as a whole, the Prediction Task, Real/Fake Task, and Mystery Mixture Task provide some interesting insights into how these TAs reasoned from a theoretical model to experimental data or from experimental data to a theoretical model. With only a few exceptions, the survey and interview participants had knowledge of the underlying probability structure in the candy jar context and appeared to be capable of attending to multiple aspects of the distribution. This is not surprising given that these TAs had all taken at least one graduate statistics course. In fact, it is surprising that there were a few TAs who did not demonstrate knowledge of the hypergeometric or binomial probability distribution. Yet, despite TAs' knowledge of formal probability distributions and their understanding of measures of center, shape, and spread, many of the survey and interview participants did not appear to access this wealth of knowledge when making decisions about the Real/Fake Task and three of the interview participants did not access this knowledge when making decisions about the Mystery Mixture Task. This is a surprising and unexpected finding.

These TAs, perhaps from lack of experience or lack of focused experience with experimental data, encountered considerable tension as they attempted to make inferences about the likelihood of certain types of outcomes in 50 trials of size 10 or

about an unknown population parameter. The source of this tension seemed to reside in their lack of knowledge of the concept of bounded variability. These TAs seemed to resolve their tensions by falling back on more simplistic ways of analyzing the experimental data in order to draw a conclusion. In the Real/Fake Task, the population proportion is known and the goal is to investigate which classes of graphs are less likely to occur as a result of the experiment. The TAs in this study largely focused on the shape of the distribution for each graph in order to make their decision. TAs appeared to focus on the unevenness in frequencies of each empirical sampling distribution on the Real/Fake Task. This expectation appeared to be a driving force behind TAs' decision-making processes. Torok and Watson (2000) observed this tendency in middle and high school students. In addition, TAs *own* images of the theoretical model shaped their conclusions.

In the Mystery Mixture Task the population proportion is unknown and the goal is to investigate what the population parameter is likely to be. The TAs in this study focused on measures of center or the shape of the graphs in making their predictions. In particular, Sandy focused on the skewed shape of the distribution, but only seemed to feel comfortable arguing that there were more yellow candies than red. Amanda and Joe used a single measure of center, the mode, and took a modal average in order to justify their predictions. Finally, Andy focused on measures of center, the median and mean, and looked at the average of these in order to justify his prediction.

While none of these methods for examining the Real/Fake or Mystery Mixture Tasks is incorrect, it is surprising that with the arsenal of statistical knowledge

available to them, for the most part they chose to only use the most simplistic of techniques. And by simplistic techniques I mean readily accessible to a layperson or someone with little to no prior statistical experience. In fact, the tendency for TAs to use shape and/or mode to justify their decisions to the Real/Fake and Mystery Mixture Task are consistent with the types of reasoning Shaughnessy et al. (2004a&b, 2005) observed in middle and secondary students. TAs who did use slightly more sophisticated techniques or incorporated multiple attributes of the distribution into their response tended to provide responses on par with the task design. For example, Andy provided the closest approximation to the Mystery Mixture simply by using the mean and median, which are more sophisticated and reliable measures of center than the mode.

The findings presented in this chapter have significant implications for TAs' statistical knowledge for teaching. The fact that these TAs experienced difficulty accessing and applying their knowledge of distributions in an experimental context suggests a limitation in TAs' ability to teach their students how to connect core concepts to an experimental context. Chapter 6 addresses the potential impact of the findings in this chapter on undergraduate statistics education. Prior to engaging in that discussion, I turn to the second significant theme related to TAs' statistical content knowledge: the different ways TAs appeared to interpret sampling and statistical inference problems.

CHAPTER 5

TAS' CONTENT KNOWLEDGE OF SAMPLING AND STATISTICAL

INFERENCE

The purpose of this chapter is to highlight the second significant theme that

emerged in my analysis of the data on TAs' statistical knowledge. This theme relates

to how the TAs in this study reasoned about sampling and statistical inference tasks,

and how they conceived of the relationship between probability and sampling, and

probability and statistical inference. In particular, the TAs in my study appeared to

reason about these tasks along a developmental spectrum, ranging from no connection

to stronger connections of a long-term relative frequency interpretation of probability.

TAs who did not connect statistical inference concepts to long-term relative

frequencies appeared to reason in a manner consistent with Konold's (1989) outcome

approach or Kahneman and Tversky's (1972) representative heuristic. In addition,

TAs' reasoning appeared to be situational and depended on their understanding of a

particular context.

This chapter is partitioned into three main sections. In Section 5.1, I briefly

describe the rationale for the two tasks discussed in this chapter, the Unusual Sample

Task and the Gallup Poll Task. I also provide an overview of two different

interpretations of probability – frequency and subjective. I discuss the implications of

each view of probability on interpreting sampling and statistical inference problems.

In Section 5.2, I provide a conceptual analysis for the Unusual Sample Task, followed by excerpts of TAs' reasoning on the sampling task that highlight the two views of the sampling context that emerged in my study. In Section 5.3, I provide a conceptual analysis for the Gallup Poll Task, followed by excerpts of TAs' reasoning on this statistical inference task that highlight two different views of confidence level that emerged from my study.

5.1 Two Different Interpretations of Probability

Both the Unusual Sample Task and the Gallup Poll Task were included in the task based interviews for TAs in this study because these tasks serve as proxies for assessing TAs' understandings of: (1) the connection between probability and sampling, and probability and statistical inference; 2) the role of sample size in sampling variability; and, (3) the image of repeatability of the sampling process (i.e., the recognition that statistical inferences are based on long-term relative frequencies). Recall that these three knowledge components were identified in my review of the research literature (see Chapter 2) as essential for a robust understanding of sampling processes and their relation to statistical inference.

There is a strong relationship between probability and inferential statistics. An illustration of the relationship between probability and inferential statistics is shown in Figure 30 (Devore, 2004, p.6). In inferential statistics, information about the properties of the population is not known, but the properties of a sample can be used to draw conclusions about the population (Devore, 2004). For example, suppose 350 voters in a sample of 500 voters gave the president a low approval rating. The true percentage

of voters in the entire population who disapprove of the president's work is unknown, but the sample is used to make inferences about the population. In probability problems, information about the properties of the population is known and that information is used to answer questions about a sample taken from that population. For example, if an experiment consists of rolling a fair die, then the question, "What is the likelihood that a six is rolled?", can be answered. This may appear to be a straightforward question, but how one interprets such a probability statement (i.e., a likelihood statement) is a key issue in this chapter. I turn now to a discussion of two different interpretations for a probability statement.

Figure 30: Relationship between probability and statistics
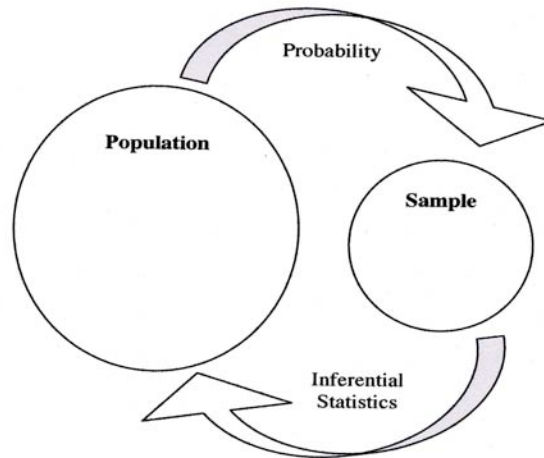
A frequency interpretation of probability is based on long-term relative frequencies. This interpretation is widely accepted by the larger statistical community and forms the basis for introductory probability and statistical inference curricula as well for more advanced study. A long-term relative frequency approach to probability

assumes an experimental situation that can be performed repeatedly, where each

repetition is performed in an identical manner and each outcome is independent of the

previous outcomes. A long-term relative frequency interpretation for rolling a fair die,

like in the example from the previous paragraph, is that if one were to roll a die under

identical conditions, a large number of times, then over the long run one could expect

the relative frequency of any given number to approach 1/6. To generalize this

perspective, consider an experiment performed *n* times and some event *A* that may

occur on some replications of the experiment. Suppose event A occurs *n(A)* times.

Then the ratio of *n(A)/n* is the relative frequency. Devore (2004) provides the

following standard interpretation of the long-term relative frequency in this situation.

> Empirical evidence, based on the results of many of these sequences of
> repeatable experiments, indicates that as *n* grows large, the relative
> frequency of *n(A)/n* stabilizes. That is, as *n* gets arbitrarily large, the
> relative frequency approaches a limiting value we refer to as the
> limiting relative frequency of the event *A*. The objective interpretation
> of probability identifies this limiting relative frequency (p. 60).

This view of probability indicates that when probabilities are applied to events, they

are meant to suggest what is likely to happen when the experiment is conducted a

large number of times. The Law of Large Numbers captures the essence of the long-

term relative frequency view of probability, identifying what is likely to happen when

an experiment is repeated a large number of times. In addition, this view of

probability, based on the long-term relative frequencies of events, also serves as the

basis of interpretation in inferential statistics. Evidence from my data analysis suggests

that some TAs maintained this perspective of probability when approaching certain tasks.

An alternative interpretation of probability is one where the experimental situation is not understood to be repeatable and probability statements are made according to an individual's degree of belief that an event will occur once and only once. This interpretation of probability is often referred to as subjective probability (Winkler, 1972). In my analysis of the data, it appears that some TAs expressed a belief of probability with reference to the *particular* individual event. Thus, these TAs' perspectives appear more similar to this subjective interpretation of probability. In this chapter, I use TAs' own interpretations of the Unusual Sample Task and Gallup Poll Task to highlight the extent to which TAs connected probability to sampling and/or statistical inference, and what, if any, evidence exists for how these TAs interpreted probability statements. Also, I show that TAs' interpretations of probability and its relation to sampling and statistical inference appeared to depend on the particular context.

5.2 The Unusual Sample Task

The Unusual Sample Task is a sampling problem. The task was given to all TAs who took the survey, and formed the basis for follow-up discussion for those TAs who participated in the interviews. The task served as a means for investigating TAs' thinking on sampling and its relation to probability. The task is shown in Figure 31.
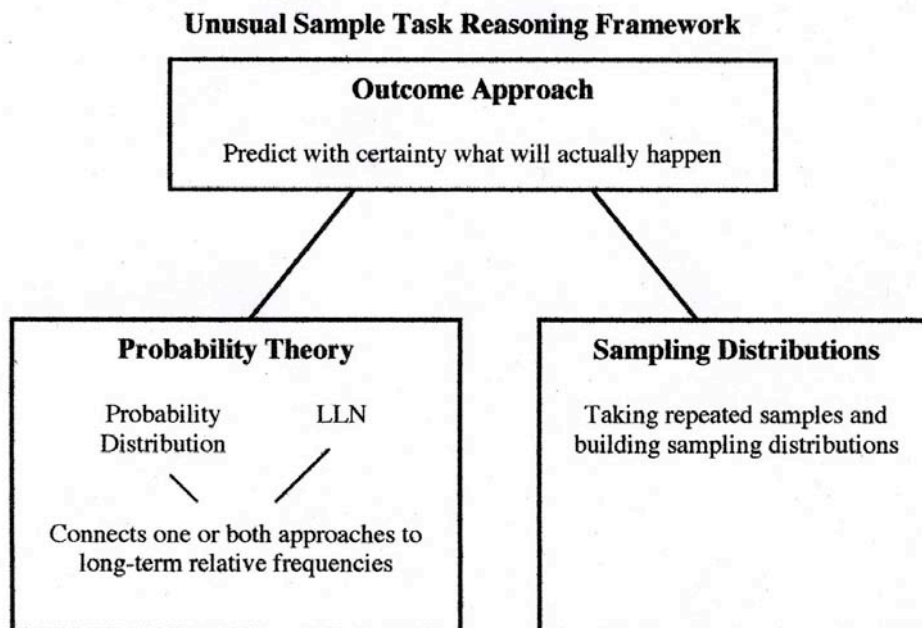
Figure 31: Unusual Sample Task

> ### The Unusual Sample Task
>
> Researchers from Erie County in Buffalo were studying the weight of Grade 5 children. The researchers went to 2 schools: one school was in the center of the city and one school was in the country. Each school had about half girls and half boys. The researchers took a random sample from each school: 50 children from the city school, 20 children from the country school. One of these samples was unusual because it had more than 80% boys. Is the unusual sample more likely to be the large sample of 50 from the city school, the small sample of 20 from the country school, or are both samples equally likely to be the unusual sample?

5.2.1 Conceptual Analysis of the Unusual Sample Task

The conceptual analysis and framework I provide here, built partially from existing research literature, developed during my analysis of the data. There are three general categories of reasoning on this task: (1) Outcome Approach (observed by Konold, 1989), (2) Probability Theory, and (3) Sampling Distribution. Figure 32 provides a visual representation of this conceptual framework. I begin with the Outcome Approach and continue discussing each approach in order.

Figure 32: Conceptual Framework for the Unusual Sample Task

Unusual Sample Task Reasoning Framework

Outcome Approach

Predict with certainty what will actually happen

Probability Theory

Probability Distribution   LLN

Connects one or both approaches to long-term relative frequencies

Sampling Distributions

Taking repeated samples and building sampling distributions

In the Outcome Approach interpretation of the problem, one views the sample of 20 from the country school and the sample of 50 from the city school as individual units – *one* sample of size 20 and *one* sample of size 50. The focus is to predict whether the *particular* sample of 20 from the country school is more or less likely to be unusual than the *particular* sample of size 50 from the city school. Such an interpretation of the problem fits with a subjective interpretation of probability because there is no image of repeating the sampling experiment and thinking about what is likely to occur over the long run; rather, the desire is to predict which of these particular samples *is* the unusual one. This interpretation is likely to lead to the conclusion that both samples are equally likely to be unusual based on the logic that if

the ratio of boys to girls in the populations are the same and the samples are random, then *any* outcome is a possibility for either sample; thus, there is no way to tell for certain. This is the type of logic observed by Konold (1989) and Konold et al. (1993). However, it could be possible that a person would reason according to the Outcome Approach and answer that the small sample is more likely to be unusual or that the large sample is more likely to be unusual. In this sense a decision for the small or large sample could be based on a 'hunch' or some probability calculation, yet the focus is still on the individual's certainty in the particular samples and determining which one *will* be unusual versus which one *is likely* to be unusual.

The Probability Theory category can be broken up into three components[23]. First, one could connect this situation to a probability distribution – in particular, a binomial probability distribution, because the assumptions of the binomial are approximately satisfied. Then the probability that the sample of 50 contains 80% boys could be computed and compared to the probability that the sample of 20 contains 80% boys. Applying the binomial probability distribution to the city school yields the following likelihood estimate that the unusual sample is the one of size 50:

$\binom{50}{40}(0.5)^{40}(0.5)^{10} \approx 0.0000912$. Applying the binomial probability distribution to the

country school yields the following chance of the unusual sample coming from the

---

[23] I use components to suggest three different subsections of reasoning that fit within the Probability Theory category and do not necessarily imply a linear ordering to these stages, although this is a possibility.

sample of 20: $\binom{20}{16}(0.5)^{16}(0.5)^{4} \approx 0.004$. While neither sample is very likely to

contain 80% boys, it is certainly *less likely* to happen in the sample of size 50.

Applying the binomial probability distribution to this sampling problem requires: (a)

an understanding that the assumptions of repeated sampling (where the experimental

unit is of size one), independence, and sampling with replacement or large population

sizes are warranted; and, (b) an ability to connect ideas of sampling to probabilities. A

TA would not necessarily have to have an interpretation of probability consistent with

a long-term relative frequency to apply this formula. In a sense, this calculation would

be a fairly routine procedure for a statistics TA, and without knowing how the TA

interpreted the information about the underlying population proportions (i.e., the

probability of obtaining a boy as being 0.5) or the probabilities obtained from the

calculations, it would be unclear whether an image of long-term relative frequencies

was unequivocally part of the TAs' mental scheme.

   A second Probability Theory perspective includes relating the Unusual Sample

Task to the Law of Large Numbers (LLN), or to the concept that as sample size

increases, sample variability or standard deviation tends to decrease. Again, citing the

LLN or this relationship between sample size and sample variability need not entail a

view of probability as a long-term relative frequency because one could simply be

recollecting how a theorem is applied without a full conception of the imagery or

reasons behind the theorem. In addition, one might have an understanding that a

sample of size one will either yield a boy or girl, and if the entire population were

sampled, the true population parameter would be obtained. Yet, a person with such an understanding might not have a sense for what happens in between these two extremes.

A third Probability Theory perspective entails an argument using a probability distribution and/or the LLN, with explicit evidence of a long-term relative frequency perspective. Here the focus is on repeating the experiment, but in this case the unit of the experiment is size 1. So the sample of size 20 is seen as 20 repetitions of size 1, and the sample of size 50 is seen as 50 repetitions of size 1. The probability distribution argument here yields the same probabilities as above (0.000092 for the sample of size 50 and 0.004 for the sample of size 20). However, the mental imagery explicitly connects to the idea of repeating the experiment and to the interpretation of the two probabilities obtained. Similarly using the LLN with an explicit connection to long-term relative frequencies suggests that as the sample size increases (i.e., as the number of trials of this experiment increases), the relative frequency of the event tends toward the true mean. This imagery leads to the conclusion that the small sample is more likely to be unusual. A visual image for this interpretation is shown in Figure 33.

Figure 33: Law of Large Numbers

n = Sample Size
As n increases the relative frequency stabilizes
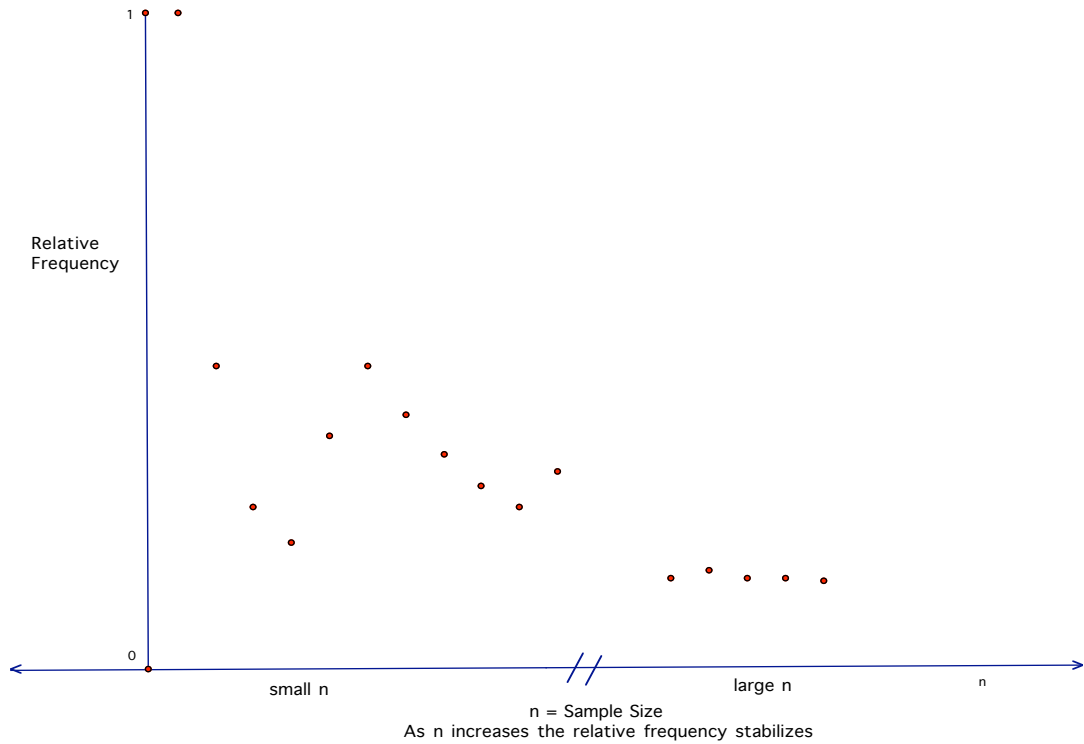
In the Sampling Distribution category, the focus is on repeating the experiment, but the unit of the experiment is a sample of size 20 and a sample of size 50. Taking repeated samples of size 20 from the country school and of size 50 from the city school, over and over again, and each time computing the value of the sample statistic – number of boys, enables the construction of the sampling distributions for the samples of size 20 and for the samples of size 50. In this case, the sample of 20 and the sample of 50 are each seen as a particular case from larger classes, each consisting of all the different sample possibilities for those sample sizes. This imagery, together with that of the sampling distributions for each sample size, also leads to the

194

conclusion that the small sample is more likely to be unusual. A visual image for this interpretation is shown in Figure 34.
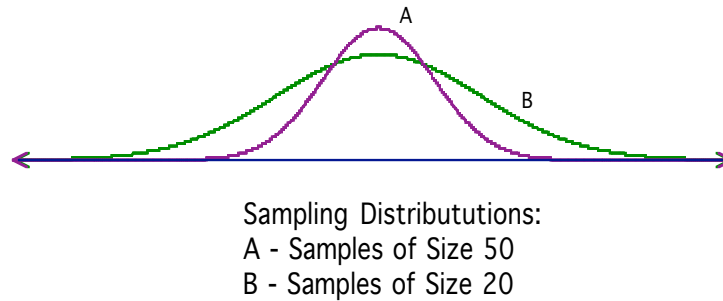
Figure 34: Distributions of sample statistics

Sampling Distribututions:
A - Samples of Size 50
B - Samples of Size 20

The sampling distribution for the samples of size 20 will be more spread out, and thus it is more likely to grab a random sample (from all those possible samples) with a sample proportion further from the center (and more unusual). This image certainly entails the idea of long-term relative frequency of probability and could also be connected to LLN.

5.2.2 TA Reasoning about the Unusual Sample Task

Table 20 shows the survey responses to the Unusual Sample Task along with a synopsis of how TAs reasoned about the task. Table 20 shows that 55 of the 68 TAs who took the survey indicated that the small sample is more likely to be unusual. In addition, 46 of those 55 TAs indicated that sample size plays an important role in sampling variability; in particular, as the sample size increases, the sampling variability decreases.

Table 20: Survey responses to the Unusual Sample Task

TAs Response	TAs' Reasoning				
	Equal ratio & Random sampling	Probability Distribution	Law of Large Numbers	Other	Totals
Small sample is more likely to be unusual	0	6 (8.8%)	46 (67.6%)	3 (4.4%)	55 (80.9%)
Both samples equally likely to be unusual	11 (16.2%)	0	0	1 (1.5%)	12 (17.6%)
Large sample is more likely to be unusual	0	0	0	1 (1.5%)	1 (1.5%)

Many TAs cited the Law of Large Numbers for their reason or provided a simple sentence such as, "as sample size increases the standard deviation decreases" or "as sample size increases the percentage of boys to girls evens out". Invoking the Law of Large Numbers or the two previous explanations is suggestive of a long-term relative frequency perspective. However, there are at least two difficulties with drawing such a conclusion. First, it is unclear if TAs merely echoed a response they've heard before. Second, it is unclear how TAs interpreted their solutions to the problem. That is, when a TA answers the question, 'which sample is *likely* to be unusual?', does their interpretation entail the aspect of certainty or likelihood from a long-term relative frequency perspective?

Twelve TAs responded that both samples were equally likely to be unusual, and one TA indicated that the large sample was more likely to be unusual. Eleven of the 12 "equally likely" responses suggested that because both school populations had an

equal ratio of boys and girls and the samples were randomly drawn, there was no reason to suspect that either sample was more likely to be unusual. The response of these 11 TAs provides some indication that these TAs did not view the sample of size 50 and the sample of size 20 as one particular case of many possible sample cases. That is, these TAs may not have interpreted this problem as being related to the long-term relative frequencies. Rather, these TAs may have been answering the question: "For these two *particular* samples, one of size 20 and one of size 50, is one more likely to be unusual than the other?".

During the first interview, I followed up on the Unusual Sample Task with the intention of gaining a clearer articulation and understanding of TAs' interpretations of the task. Table 21 shows the survey responses to the Unusual Sample Task for the five TAs I interviewed.

Table 21: Interview Responses to the Unusual Sample Task

TA	Response	Reason
Amanda	Both samples equally likely to be unusual	"We are told that both schools have approximately the same percentage of boys and girls, so neither is more likely to have produced the unusual sample."
Sandy	Both samples equally likely to be unusual	"Each school had about half girls and half boys."
Joe	Small sample more likely to be unusual	"Higher sample sizes are less likely to show weird things when taking a random sample than smaller ones. $(.5)^{20}$ is much larger than $(.5)^{50}$, for example."
Andy	Small sample more likely to be unusual	"Law of Large Numbers." (No explanation)
Sam	Small sample more likely to be unusual	"It has a small sample size. If I choose 2% of the students randomly, the probability that they will all be boys will be higher, as compared to choosing a higher percentage."

Joe, Andy, and Sam indicated that the small sample was more likely to be unusual. Joe suggests that larger samples are less likely to "show weird things", and Sam seemed to simply state that a smaller sample size is more likely to produce all boys. Implicit in Joe's and Sam's suggestions could be the recognition that as the sample size increases the sampling variability decreases, a major consequence of the LLN. Yet, this understanding is not explicit in Joe's and Sam's justifications. Andy merely cited the LLN as his justification. In addition, Joe provided what appears to be a binomial probability distribution argument. Amanda and Sandy indicated that both samples were equally likely to be unusual because of the equal ratio of boys to girls in the population. The type of responses provided by these TAs makes it difficult to draw any conclusions about how TAs actually interpret a probability statement, or in what ways they relate this sampling problem to probability. It is also difficult to draw any conclusions about their having any explicit sense of variability.

At the beginning of the interview I asked each of these TAs to explain how they interpreted the problem and to elaborate on the reasons for their choices. I believe that the follow-up questions and subsequent TA responses provide evidence that Sandy and Amanda reasoned about this task from an Outcome Approach, and that Andy, Joe, and Sam reasoned about this task from Probability Theory. There is some evidence that Andy and Sam reasoned from a Probability Theory perspective with long-term relative frequency imagery. There is no evidence that any of these TAs reasoned with a Sampling Distribution approach, and it is difficult to infer from the data whether or not the long-term relative frequency perspective was explicit in Joe's mental scheme.

First, I examine the responses of Andy, Joe, and Sam and follow with the responses of Amanda and Sandy. I group the interview responses in this way because of the commonalities in the TAs' responses.

TA reasoning: Andy, Joe, and Sam

Andy's interview responses to the Unusual Sample Task suggest that he reasoned about this problem in terms of long-term relative frequencies of events. In the first exchange, Andy referenced the extreme cases in his discussion of the problem. The following excerpt provides Andy's initial explanation of the problem during the interview.

> Interview 1: Unusual Sample Task:
>
> > **Interviewer:** So my first question for you is to walk me through how you interpreted the problem and how you went about solving it.
> > **Andy:** Okay, well I started with the base. I started with the extreme cases and said suppose there was only one child in the sample. You know if the small sample contains a single child then its either going to be 0% boys or 100% boys. The small samples are going to have these extremes. If you take a large sample where there is 50% boys and 50% girls in the general population, the larger sample you get, the closer you are going to get to 50%. Well that's the Law of Large Numbers. You are going to tend toward the average of the population.

Andy seems to connect this problem to the Law of Large Numbers and his understanding that large samples *tend* to be less variable than small samples. Andy's discussion of the extreme case (highlighted in the above excerpt), where a sample of size one is collected, suggests that he at least has some image of the ends of this long-term relative frequency process. His utterance that "the larger sample you get, the closer you are going to get to 50%", together with his use of the word "tend", is also

an indication that he conceptualizes the entire process where the limiting relative

frequency value points to the population parameter. When I pushed Andy about his

use of the LLN, indicating that some TAs did not think an application of the LLN was

appropriate for the Unusual Sample Task, Andy continued to argue from this

perspective.

Interview 1: Unusual Sample Task:

> **Interviewer:** There were two main things that came out on this survey. People responding exactly as you did. They didn't necessarily cite the Law of Large Numbers, but they walked me through something similar to what you did. And then there were other TAs that responded that both samples were equally likely. So that both samples would be equally likely, and one response for this that appeared often was that the samples were randomly chosen and that the difference between the sample sizes 20 and 50 wasn't significantly large enough to matter. How might you respond to that TA? Do you agree or disagree?
> **Andy:** Well, I'd have to disagree because again you go to the extreme. It's not significant between a 50 and a 20 – well one's more than twice the other, that's going to *tend* [***Andy's emphasis***] to be significant…. It's not about sampling the entire, exhaustively sampling the population. It's about taking a sub-sample. The larger the sub-sample, the more likely you are to get the true average, the population average. The smaller the sample the less likely, you *tend* [***Andy's emphasis***] to get extremes.

Here Andy's utterances suggest looking at different sub-samples from the population

and the process of taking larger sub-samples as a means for approaching the

population average. Further, Andy seemed to emphasize the words "tend" and "more

likely" during the discussion. His language is suggestive of the long-term relative

frequency process although he does not use those words explicitly.

Sam did not articulate his position as clearly as Andy, but the following excerpt

also suggests that Sam viewed the Unusual Sample Task with an understanding of the

consequence of the LLN and a long-term relative frequency perspective.

Interview 1: Unusual Sample Task:

> **Sam:** … Say we have 20,000 PSU students and I collect the heights of every student. Then I know for sure the average height. If I can collect a sample of 10,000 [*heights*] I will be closer to collecting all heights than if I only collect a sample of 100 [*heights*]. So this larger sample will give a better approximation of mean height [*of all PSU students*]. If I only collect one sample [*a sample of size one*], I just get one height and it may not be very accurate.

Like Andy, Sam referenced the extreme cases, a sample of size one, and compared

that to the entire population. Sam also mentioned what might happen in between

during this process. He suggests that a sample of 10,000 is closer to the entire

population than a sample of size 100 and that as this sample size increases, the larger

sample will tend to be more accurate. When I pressed Sam about his reasoning, and

suggested that some TAs argued that the difference in sample size was not large

enough to matter, he did not appear comfortable arguing his case. In addition, Sam's

utterance that a "larger sample will give a better approximation of the mean height" is

also suggestive of the Outcome Approach perspective in that Sam uses the words *will*

give rather than *is likely* to give. It is difficult from the data to assess how Sam

intended the words *will give*. In the same manner it is difficult to assess from the data

how Andy intended the phrases *tends to* or *is likely*. Phrasing that is often used to

discuss probability or limiting values, such as *likely* or *tends*, are used in a plethora of

non-stochastic ways in the English language; thus in hindsight I should have probed

these TAs about what they meant by those phrases.

Joe's survey response, "Higher sample sizes are less likely to show weird things

when taking a random sample than smaller ones. $(.5)^{20}$ is much larger than $(.5)^{50}$, for

example" suggests that he was using both the LLN and a probability distribution

argument. During the interview, Joe referenced the extreme cases. The following

exchange is Joe's initial explanation for how he reasoned about the Unusual Sample

Task.

Interview 1: Unusual Sample Task:

Interviewer: Talk me through how you thought about this problem and how you went about solving it.
Joe: Sure, I mean the general rule is, as the sample size increases your distribution curve approaches the population. If your sample is as big as the population it's going to be the population.…So if the population is 50/50 boys and girls, as the number of boys and girls we pick goes up, the closer we'll get to 50/50. But if we just pick one kid, it's either a boy or a girl and we have 100% boys or 100% girls.
Interviewer: Okay, you had this .5 raised to the 20th power is larger than .5 raised to the 50th power on your survey. So I'm kind of wondering how these calculations, how does that show what you just explained to me?
Joe: Yeah. This is just a sense-making thing. I mean that would kind of be the chance of having all girls or all boys in a sample of 20. I think. I'd have to, I mean yeah, right?…Well, we had a 50/50 chance when we were pulling our sample from the population. So if you just pull one (*child*) it's 50%. The chance of getting two girls is 25%, .5 times .5. The chance of getting 3 girls is 1.25. I'm sorry .125. So the chance of getting 20 girls is .5 raised to the 20. That was a sense-making thing. Even though this is an exceptionally small number [*.5 raised to the 20*] it is still much larger than this number [*.5 raised to the 50*].

In the first highlighted utterance above, Joe discussed the extreme case of picking only

one student and that the student would have to be either a boy or a girl. He also

discussed the end of the process whereby the sample consists of the entire population and thus, the population parameter is obtained. However, Joe does not articulate what happens in between these two extremes. There is no explicit evidence in his initial utterances that he has an image of repeatability consistent with a long-term relative frequency perspective. Also, his first utterance, where Joe says, "the general rule is…" suggests that he may merely have invoked a formula or a procedural process.

As I asked Joe more about the computations he provided on his survey, his response suggests that he related the Unusual Sample Task to a binomial probability distribution problem (although he did not explicitly use this terminology) in which he mentally estimated the probability of getting 100% boys or girls in a sample of 20, versus 100% boys or girls in a sample of 50. The second highlighted portion of the excerpt above implies that Joe thought about the sample of size 20 as 20 trials and compared that to what was likely to happen in 50 trials. To each trial, he assigned the probability of success as 0.5 and made an implicit assumption that each trial was independent so that he could multiply the probabilities together. For instance, when Joe says the chance of getting two girls is 0.5 times 0.5, or 25%, he was treating the sampling situation as approximately binomial, where the probability of success is 0.5. By looking at the probability of getting 100% boys in the sample of size 20 versus the sample of size 50, Joe was able to infer that the likelihood of getting a sample of 80% boys is likely to be larger in the small sample of size 20. However, his binomial probability distribution approach does not provide any indication of how he interpreted the probabilities obtained from his calculations. In addition, his response

that smaller samples tend to be "weird" may be the result of his application of the formula, rather than a deep understanding of the LLN and an image of long-term relative frequencies.

When I pressed Joe about his decision and suggested that some TAs responded that both samples were equally likely, he appeared unable to provide a clearly articulated justification for his choice.

Interview 1: Unusual Sample Task:

> **Interviewer:** So when I think back to the survey and how TAs responded, they responded pretty much how you just did or this way, that both samples would be equally likely to be the unusual one. And one person reasoned that the samples were randomly chosen and that the difference between the sample sizes 20 and 50 wasn't significantly large enough to matter. How might you respond to that TA? Do you agree or disagree?
> **Joe:** I don't know enough about the math to know if 20 and 50 are close enough. And I don't know what enough would be here. Yeah, I don't know what kind of null hypothesis there would be for that. So I'd want to do some math before disagreeing, but my intuition is that it's more likely to come from the smaller sample size, given that the population in the schools are both 50/50 boys and girls. If I'm taking 50 from one and 20 from the other, the 20 is more likely to be weird than the 50. But I don't know if it's statistically significant there.

Joe had strong intuitions for this problem. In the highlighted utterance above he says that his intuition tells him the smaller sample size is more likely to be unusual. Yet, he could not provide a clear articulation of the process by which the values of the statistic for large samples converge toward the population parameter. Without a more robust understanding of the LLN and a relative frequency perspective of probability, he seemed unsure about how to disagree when I played devil's advocate.

In sum, Andy, Joe, and Sam all used a Probability Approach to reason about the Unusual Sample Task, but there were distinctions and differences within that approach. Andy and Sam appealed more to the LLN suggesting that as sample size increases the sample estimate tends toward the population parameter. There is some evidence from Andy's and Sam's utterances that they had a long-term relative frequency interpretation of the process. Joe appealed more to the underlying binomial probability structure to reason and there was no evidence in Joe's utterances that he had a mental image of repeatability or long-term relative frequencies. It would have been beneficial to have asked Andy, Sam, and Joe what they meant by "likely" or "tends to" because such questioning may have provided stronger evidence of how each of these TAs interpreted probability statements.

TA reasoning: Amanda and Sandy

Amanda's and Sandy's interpretations of the Unusual Sample Task stand in stark contrast with the interpretations provided by Andy, Joe, and Sam. Both Amanda and Sandy appear to perceive the situation from a single outcome perspective, although perhaps to differing degrees. Amanda's survey response suggests that she did not think either sample was more likely to be the unusual sample, and although this remained her bottom line during the interview, it also became apparent that she considered the role of sample size in her response.

Interview 1: Unusual Sample Task:

Interviewer: Take a minute to read over the first question from the survey. Your response is here. So reread the problem and your response, and then

walk me through how you thought about the problem and how you solved it.

…

Amanda: I feel like I cannot tell. There's certainly a lot of different ideas that float around in my head, but in the end I keep coming back to that we were told in the general populations we had 50% boys, 50% girls or close enough. So I've considered things like since it's the extra information we have to go on, the sample size, would it be less likely the large sample I take the more representative it's going to be of the general population. So would it be less likely that the 50 students would be more closely split 50/50 than the 20 students [*she has this backwards here, but I think she is just misspeaking as later she goes on to say it correctly*]. But I don't feel confident enough to say that there's enough information there for me to say that one of them is more likely to have produced an odd sample. Because I don't think 50 is that large. I know in our textbooks we talk about 30 being a large sample, but I don't feel like 50 is very big.

Interviewer: So kind of what I'm hearing you saying is that 50/20 is not enough to make you feel confident that you could say. So if this had been 100/20 would that be different for you in terms of answering this?

Amanda: If there was a dramatic difference and the 50 was larger, say 100. I might lean more heavily on the assumption that the larger the sample I take the more it's going to mimic the distribution it came from and I would be able to say that it's probably closer to 50/50 than the 20 sample would be.

On the survey, Amanda responded that both samples were equally likely to be the unusual sample because the ratio of both populations was roughly 50/50, and the samples were randomly drawn. However, in this excerpt we see that Amanda also considered the role that sample size plays in sampling variability. Yet, because Amanda did not think the difference in sample sizes was significant, she returned to the fact that the ratio of boys to girls was 50/50. For Amanda, the difference in size between the two samples would need to be much larger in order for her to believe that the large sample would be more representative of the population. In addition, when I raised an alternative response to the problem and suggested the Law of Large

Numbers, Amanda did not appear surprised by the use of the LLN in this context. She also recognized that the situation could be examined with a binomial probability distribution. The following exchange took place after I brought up the LLN.

Interview 1: Unusual Sample Task:

> **Interviewer:** Okay, so let me tell you first that pretty much most responses fell between the small sample as the unusual one, or both samples are equally likely to be the unusual one. Some folks that said the small sample would be more likely to be the unusual one and used the Law of Large Numbers. What would you say to a person who argued the opposite of you, and said I think the unusual sample is from the sample of 20 because of the law of large numbers?
> **Amanda:** [*Laughs*]. The Law of Large Numbers. Because of the Law of Large Numbers?
> **Interviewer:** Yeah.
> **Amanda:** Okay, so the weak Law of Large Numbers says that a value is going to center around… that it's going to converge to its mean, right?
> **Interviewer:** Yeah.
> **Amanda:** And I could possibly be way off on this, but large numbers implies to me that this is occurring as *n* gets larger. It makes me completely doubt myself when you say well somebody else said blady blah. But it's my understanding that when we are talking about the Law of Large Numbers it's going to be true as the sample size grows. …Wait, wait I'm putting it all together. I'm putting all the pieces together. So the person would be saying the 20 is more likely to be the unusual one because of the Law of Large Numbers, because as my sample size grows I have to get focused in around that average value. Which is what I'm saying. Did I say it backwards? That the 50 would be less likely because it's a larger sample its going to be more representative of the general population versus the 20 being a smaller sample is not necessarily as accurate a representation of the population. So the 20…[*laughs*] I really hope I didn't just say all of that backwards. So the 20, if I had to pick one, it would be 20 to be more likely to produce the off sample. Just because it's smaller than 50 and as our sample size grows we're going to narrow in on that 50/50 girls/boys. But 50 in my mind, and this is not based on anything other than comfort level, does not strike me as large enough to say, 'yes in this situation 50 is more likely to be the good sample versus 20 the bad sample'.

Initially Amanda appeared nervous and doubted her response when I indicated that some TAs concluded the small sample was more likely to be the unusual one because of the Law of Large Numbers. However, Amanda seemed inclined to agree with these other TAs that the large sample would tend to be more representative of the original population, although she did not seem to feel comfortable with this conclusion because the difference in sample sizes was not significant enough *for* her.

Amanda also recognized the underlying probability structure, but she did not appear to use it as she worked through the problem.

Interview 1: Unusual Sample Task:

Interviewer: Okay, is there anything in stats that allows you to determine something like this, like if that would be a large enough difference to actually answer a question like this with determination?
Amanda: [**Long pause**]. I believe so. I think I spent a lot of time on this question trying to think if there was a way that I could calculate to make me feel good about my answer. But certainly you could break this down into a binomial problem where you're looking at your sample and you're told the population is about half and half. So if you're looking at boys and girls you know the probability of getting boys and girls. So of your sample of 20, what's the probability that you're going to get more than 80% boys? Versus if you had a sample size of 50, what's the probability that you're going to get more than 80% boys? And you could do a comparison of how those probabilities turn out. I mean clearly they're not going to be equal. I imagine not because the differences in sample sizes, but without sitting down and performing a test like that I'm not willing to say that one is more likely than another.

Amanda discussed the underlying binomial probability structure of the problem and appeared to be able to view the problem from a number of perspectives. In this last exchange she showed flexibility by looking at the sample of 20 as 20 trials and the sample of 50 as 50 trials, where each trial has a 50% chance of producing a boy and a

50% chance of producing a girl. Yet, her main focus appeared to be on the difference in sample sizes between the two samples, and her ability to know *for certain* if one sample versus another sample would be more likely to be unusual. The bottom line is that without an exact calculation, Amanda rested on her gut level intuition and her desire to answer, 'which sample *will* be unusual?', rather than 'which sample is *likely*, in the probabilistic sense, to be unusual?'. It appeared that Amanda wanted to be able to make her choice with *absolute certainty* rather than *probabilistically* with likelihoods.

Like Amanda, Sandy also believed that both samples were equally likely to be the unusual one, and Sandy seemed intent on answering the question of which sample *will* be unusual, rather than which sample is *likely* to be unusual. Yet, whereas Amanda seemed comfortable looking at the Unusual Sample Task from a number of perspectives, Sandy appeared to compartmentalize her knowledge of probability and sampling. She did not see a connection between the Unusual Sample Task and the underlying probability structure or its relation to the Law of Large Numbers. The following exchange provides some insight into Sandy's perspective of the Unusual Sample Task.

Interview 1: Unusual Sample Task:

> **Interviewer:** So here is the first problem from the survey. Here is the problem along with your response. Will you walk me through how you interpreted the problem and how you solved it?
> **Sandy:** We are told that we got a sample and one of the samples had 80% boys. But we don't know, we have no information how the population of those two schools are. If I would have been told that the population, you

know something in the population I would have maybe changed my answer.

Interviewer: So what is it about the population? Is it like a specific size for a population or…

Sandy: No, no, no size. You know like a school has more boys in general than girls.

Interviewer: But it says each school had about half girls and half boys.

Sandy: Yeah, exactly so that's it. That was my reason.

Interviewer: Okay that's what you were putting down here [*pointing to Sandy's explanation*] because each school had about 50/50. Why would it be different?

Sandy: Yeah, Yeah. That's how I see the problem. Because you collect a random sample. It's random definitely you can have all kinds of situations. You could have more boys than girls, and equal number of girls and boys or more girls than boys.

Interviewer: So what if you were told that the city school had more boys?

Sandy: Then I would have thought that the sample with more boys comes from the city school because I was told that it had more boys. Because you know that sometimes there were schools that were only for boys or only for girls and at some point they changed. And said ah, in the school of boys we are going to start accepting girls, so definitely in the years beginning of that process they would have more boys than girls.

This exchange suggests that Sandy was intent on determining whether the particular sample of 20 or the particular sample of 50 *was* unusual. In the first highlighted utterance, Sandy focused on the fact that the samples from each school were randomly selected making all kinds of outcomes possible; thus, for Sandy it would be impossible to determine for the *particular* samples drawn which *would be* unusual. In this utterance Sandy appeared to provide an acknowledgement of her expectations for variability in this situation and she did not appear to place any boundaries on that variation. From my interpretation, Sandy's understanding was that since only one random sample was collected there was no reason to believe that one sample versus another, no matter the difference in sample size, was more likely to be unusual. This

view is consistent with Konold's (1989) Outcome Approach in that Sandy appeared to

be answering the question, 'Which sample *will* be unusual'?

    When I brought up a TA response that differed from hers, she discredited it

quickly, as seen in the following exchange.

    Interview 1: Unusual Sample Task:

> **Interviewer:** Some of the TAs taking this survey answered exactly how you did, but some answered that the small sample from the country school…
> **Sandy:** No I see no reason for that.
> **Interviewer:** Well, one TA used the Law of Large Numbers for his reason.
> **Sandy:** The Law of Large Numbers.... I don't really see that. I don't see the relationship because this problem is about collecting a sample, you know.
> **Interviewer:** Okay, so if we are flipping a coin. Let's look at the flipping the coin example. If we're flipping a coin 20 times versus 100 times, is it more likely that we'll get half and half 20 of the times?
> **Sandy:** If you, if you flip like this, you said how many? 20 times.
> **Interviewer:** Yeah.
> **Sandy:** Versus 100 times. Here [***pointing to the 100 times written on a sheet of paper***] the probability will be closer to 0.5 than here [***pointing to the 20 flips written on a sheet of paper***].…So here the probabilities are much more closer. The more times you flip the coin the closer the probability of heads to tails comes to 0.5. Yes, I remember now. I read at some point an example. It was some guy who was in prison and he flipped a coin, I don't know, I believe 5000 times or something like that.…Look he flipped so many times and it was the probability and so on and the students were able to see how the probability approaches 0.5 the more number of times you flip the coin.
> **Interviewer:** But you don't think that's the same here because you have 50/50 boys/girls?
> **Sandy:** Yes, yes you have 50/50 boys/girls, but it's not a problem about umm. …It's not a similar problem. I don't think. Not from my point of view. You just go and you sample some people randomly, randomly.
> **Interviewer:** So flipping a coin, are you relating that more to probability versus this as being a random sampling?
> **Sandy:** Yeah. I don't know, but I'm thinking like that.… Because like here [***referring to the coin problem***] you are doing the same thing over and over again. While here [***referring to the sampling problem of boys and girls***]

you go and take a sample, a random sample. You're not supposed to go again and go again. If you get 80% boys you get 80% boys. Maybe if I go like 20,000 times to get my sample to be close to 50% boys, maybe I will get it, I don't know. …Yeah that's how I see it. Because if I go and I sample once, how it's normally to be done. You want to be unbiased, am I right? You have to be unbiased, so you go and you sample. It's not ah, umm a question about do you like your sample or not because otherwise you'll be biased and you won't be able to have a correct test done. [*She's referring to the fact that if we related the sampling situation to flipping the coin we'd have to sample many times before the probability would approach the 50/50 mark, but we do not sample many times in reality and we cannot sample many times until we get a sample that suits us because this would bias our results*]. …I was told that each school had half girls and half boys. I have no reason, if I pick 20 randomly or if I pick 100 randomly, I have no reason to believe that I will approach 80% boys more in one versus another.

I believe the previous exchange reveals that Sandy viewed the Unusual Sample Task as a sampling problem and did not see a connection between probability and sampling in this context. In the first highlighted utterance, Sandy suggested that the Unusual Sample Task was a sampling task, unrelated to the concept of the Law of Large Numbers. When we discussed a coin-flipping example, she saw the applicability to the Law of Large Numbers, but she did not see any connection between the Unusual Sample Task and the coin-flipping example because she did not think of the sample of 20 as 20 trials. Sandy did not think it was appropriate to relate the Unusual Sample Task to an underlying probability structure, nor did she think it was appropriate to think about repeating the sampling process. In fact, Sandy was adamant that the coin flipping example and the Unusual Sample Task were distinctly different because in the coin flipping example you needed to repeat the flip over and over again in order for the relative frequency to become close to the population ratio, and in the Unusual

212

Sample Task there was no possibility for repeating the sampling process (nor was there an image of hypothetical repeated sampling) (see the fifth and sixth highlighted utterances in the previous excerpts). That is to say, it appeared that Sandy wanted to predict the outcome of the particular samples drawn in this experiment and thought about the sample of 20 as its own unit and the sample of 50 as its own unit, not as independent trials. Sandy's image of the Unusual Sample Task did not appear to include hypothetically repeating the experiment and examining the long-term relative frequencies. Rather, her conceptual image included the concept of a random sample, her knowledge that both populations contained 50% boys and 50% girls, and her understanding that she was being asked to predict what *would* happen after collecting one sample of size 20 and one sample of size 50. From this perspective, Sandy argued that it was not possible to know which sample would be more likely to be unusual. Sandy's response provides strong evidence that her interpretation of the Unusual Sample Task was from a subjective probability interpretation, a measure of her degree of belief for what outcomes would occur in each sample.

Summary of TA Reasoning about the Unusual Sample Task

In sum, about 81% of TAs on the survey and three out of five TAs I interviewed used a Probability Theory approach to argue that the small sample is more likely to be unusual. Yet, there was not enough information available in the data to conclude much about how TAs interpreted the probability statement. However, what is particularly compelling is that there is strong evidence that two of the five TAs I interviewed did

not have robust knowledge of the relationship between sample size and sampling variability, and reasoned about the Unusual Sample Task using the Outcome Approach. The utterances provided by Amanda and Sandy certainly suggest that their responses were based on the degree to which they believed they could predict the outcome for this one unique event. Given that the samples were randomly selected and the ratio of boys to girls at each school was the same, Amanda and Sandy argued that there would be no way to tell which sample was more likely to be unusual. In prior studies (Kahneman & Tversky, 1972; Konold, 1989; Watson & Moritz, 2000) using this type of sampling task, researchers found that students often use the Outcome Approach, and respond that both samples are equally likely to be unusual. These studies have involved K-12 through tertiary students. Also, these prior studies indicated that students who reasoned that both samples were equally likely to be unusual did so because they were trying to say something about the particular samples collected in the study, and did not think about the concept of repeated sampling. This study provides initial empirical evidence that some graduate TAs in statistics also reason about sampling problems using the Outcome Approach.

5.3 The Gallup Poll Task

The primary purpose for including the Gallup Poll Task in this study was to investigate TAs' understanding of confidence intervals and the ways in which they may or may not connect ideas of probability and repeated sampling to statistical inference. As a place to begin an investigation of TAs' understanding of confidence

214

intervals, I initially asked them to provide an interpretation of margin of error in the following task (see Figure 35).

Figure 35: Gallup Poll Task

> **Gallup Poll Task**
>
> **Your statistics class was discussing a Gallup poll of 500 Oregon voters' opinions regarding the creation of a state sales tax. The poll stated, "...the survey showed that 36% of Oregon voters think a state sales tax is necessary to overcome budget problems". The poll had a margin of error of ± 4%. Discuss the meaning of margin of error in this context.**

The Gallup Poll Task provided an opportunity for gathering information on TAs' thinking around confidence intervals. After TAs provided their own interpretation of margin of error and confidence interval in the Gallup Poll Task, they were shown hypothetical student interpretations (Students A through F, see Figure 36) and asked to comment on the reasonableness of each interpretation.

Figure 36: Hypothetical Student Interpretations

> **Gallup Poll Task: Hypothetical Student Interpretations of Margin of Error**
> **Student A says:** The margin of error being 4% means that between 32% and 40% of all Oregon voters believe an income tax is necessary.
> **Student B says:** We don't know if the interval 32% to 40% contains the true percentage of voters that believe an income tax is necessary, but if we sample 100 times, about 94 of those times the interval would capture the true percentage of voters.
> **Student C says:** The interval 32% to 40% will be off about 4% of the time, or 4 out of 100 times.
> **Student D says:** If you performed repeated samples of 500 voters, the proportion of voters in favor of sales tax in these samples would fall within the interval 32% to 40%, the majority of the time.
> **Student E says:** I can be 95% sure that all the sample statistics will fall within ± 4% of the unknown population parameter.
> **Student F says:** The interval 36% ± 4% has a high probability (approximately 95%) of being repeated if the sample was repeated.

Many of the hypothetical student interpretations I developed were based on Liu's (2004) conceptual framework for investigating teachers' conceptions of margin of error in a polling context. The hypothetical student responses served two main functions. First, they provided an additional opportunity to gather information about how TAs thought about confidence intervals. Second, they provided an opportunity to gather information on TAs' knowledge of content and students (this will be discussed in Chapter 6). In addition to the hypothetical student interpretations of margin of error, I asked TAs to discuss the confidence level for the Gallup Poll Task and I provided two hypothetical student interpretations of confidence level and asked TAs to respond to each of these interpretations (see Figure 37).

Figure 37: Hypothetical Student Responses to Confidence level

> **Gallup Poll Task: Investigating Confidence Level**
>
> **Hypothetical Student 1:** A 95% confidence level means that you can be 95% confident that the particular interval found in the survey captures the population proportion. Do you agree or disagree with this student's interpretation of confidence? Explain.
> **Hypothetical Student 2:** A 95% confidence level means that you are 95% confident in the estimation process. That is, 95% of the time you get good interval estimates that capture the population proportion. Do you agree or disagree with this student's interpretation of confidence? Explain.
> What would the confidence level be for this Gallup poll?
> **How do you interpret confidence level in this context?**

In the sections that follow, I provide a conceptual analysis for understanding confidence intervals from two different perspectives – a frequency perspective and a subjective perspective. Following the conceptual analysis, I discuss in more detail the

different hypothetical student responses. Finally I discuss how the five TAs I interviewed thought about the Gallup Poll Task and how they responded to the different hypothetical student interpretations.

5.3.1 Conceptual Analysis of Confidence Intervals

Confidence intervals are an important component of inferential statistics because a point estimate alone does not provide any indication of how close the estimate might be to the population parameter. A confidence interval, then, is an interval estimate together with an associated measure of reliability. From a frequency interpretation of probability, a robust appreciation of confidence intervals requires an image of repeating the experiment over and over again, and thinking about the long-term relative frequency of the number of interval estimates that would capture the population parameter. Consider a random experiment with n independent repetitions from a finite population with mean μ and standard deviation σ. Then the values X_1, $X_2, \ldots, X_n$ of the random variable X represent the various means from the n repeated trials. Applying the Central Limit Theorem, as n increases, the sampling distribution of means will be approximately normal with expected value μ and standard deviation $\sigma/\sqrt{n}$. Standardizing $\bar{X}$, the mean of the sampling distribution, gives $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ which

has a standard normal distribution with an area under the curve of 0.95 within ± 1.96 standard deviations of the mean. This yields the following probability statement:

$$P(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96) = 0.95 \qquad (1)$$

Solving the inequality in (1) for μ we get the following standard formula for confidence intervals:

$$P(\bar{X} - 1.96\,\sigma/\sqrt{n} < \mu < \bar{X} + 1.96\,\sigma/\sqrt{n}) = 0.95 \qquad (2)$$

The probability that the random interval in (2) contains μ is 0.95 because at this point the problem is strictly a probability problem and the $\bar{X}$ is a random variable that has not been realized in the experiment. The problem becomes statistical in nature once a random sample is collected and the value of $\bar{X}$ becomes known. At this point, the interval has known endpoints, and it is not acceptable to say with 95% confidence that the particular interval calculated contains μ, even though that might be tempting to conclude (Devore, 2004; Hogg & Craig, 1995). The population parameter, μ, is a *constant* and the particular interval computed either includes μ or it does not. Thus, the interpretation of a 95% confidence interval

> relies on the long-run relative frequency interpretation of probability: To say that an event *A* has probability 0.95 is to say that if the experiment on which A is defined is performed over and over again, in the long run *A* will occur 95% of the time (Devore, 2004, p. 285).

That is, if the sampling process is repeated many times, one can expect that 95% of the resultant confidence intervals would include μ. Figure 38 shows a visualization of the repeated sampling concept in relation to the creation of a 95% confidence interval (Devore, 2004, p. 285).

Figure 38: Repeated sampling and statistical inference
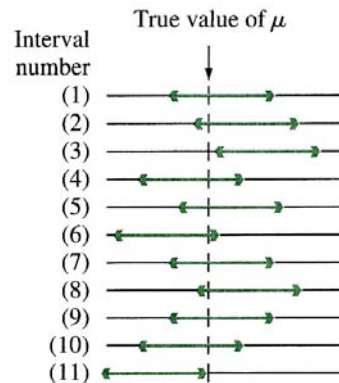
A common alternative interpretation of a 95% confidence interval is that the

particular interval calculated from the sample has a 95% chance of containing μ.

From a mathematical perspective, this interpretation has a different meaning than the

frequency perspective in that the level of confidence is in the specific interval, rather

than in the method by which confidence intervals are produced. This interpretation

would be consistent with a subjective interpretation of probability. In the statistical

community, confidence intervals with a subjective (also known as Bayesian)

interpretation are called credible intervals (Winkler, 1972). According to Winkler, in

the subjective interpretation the statistician makes a probabilistic judgment about the

population parameter, whereas in the frequency interpretation the statistician makes a

probabilistic judgment about the sample statistic (p. 394).

5.3.2 Hypothetical Student Interpretations

As noted at the beginning of this section, after TAs provided their own

interpretations to the Gallup Poll Task, I provided them several alternative

hypothetical student interpretations. I discuss each of them in more detail here, beginning with the coherent interpretations.
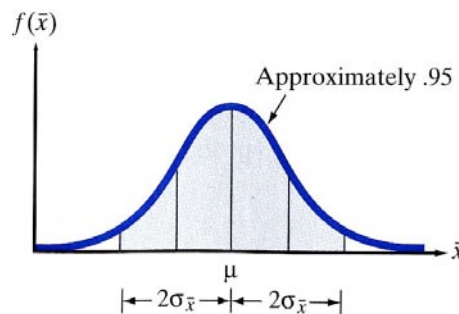
Coherent Hypothetical Student Interpretations

Liu (2004) examined an accepted textbook interpretation of confidence interval and the possible ways it could be recast to either an equivalent coherent interpretation based on the image of repeated sampling or to alternative interpretations. Recall that a standard coherent interpretation of confidence interval says: if we repeat the sampling process a large number of times, then *Y%* (where *Y* is typically 90, 95, or 99%) of the intervals, *sample statistic* ± *margin of error*, will contain the true population parameter. Hypothetical Student B's interpretation, "we don't know if the interval 32% to 40% contains the true percentage of voters that believe a sales tax is necessary, but if we sampled 100 times, about 94 of those times the interval would capture the true percentage of voters", typifies this textbook interpretation of confidence interval in the context of the Gallup Poll Task.

An equivalent definition used by Liu (2004) in her teaching experiment states that the interval, *population parameter* (*p*) ± *margin of error*, captures *x%* of the sample statistics ($\hat{p}$). Hypothetical Student E's interpretation of the poll's margin of error, "I can be 95% sure that all the sample statistics will fall within ± 4% of the unknown population parameter", is representative of this alternative construal. Notice that the standard text interpretation, $\hat{p} - E \le p \le \hat{p} + E$ (where *p* is the true population parameter, $\hat{p}$ is a sample statistic, and *E* is the margin of error), is algebraically

equivalent to: $p - E \leq \hat{p} \leq p + E$. In many respects, thinking about the percentage of

sample statistics that are within a certain distance of the unknown population

parameter is a more natural way to approach an interpretation of confidence intervals

with the conceptual entailment of the sampling distribution. The image in this

approach is to take a sample and compute its sample statistic, take another sample and

compute the new sample statistic, and do this over and over again, until a large subset

of the sampling distribution (the distribution of all possible samples of size n) has been

collected. The mean of this sampling distribution is approximately equal to the

population parameter, μ, because of the natural properties of sampling distributions.

Further, for large n, the sampling distribution is approximately normal and the interval

around μ, which contains 95% of the sample statistics, can be examined. See Figure

39 (taken from McClave & Sincich, 2000, p. 280).

Figure 39: Sampling Distribution of *x*-bar

(where *x*-bar represent sample statistics)

In addition, I added Hypothetical Student 2's interpretation of confidence level as

a coherent interpretation, which places the confidence within the method, identifying

the random error associated with the sampling process. Student 2's interpretation says, "A 95% confidence level means that you are 95% confident in the estimation process. That is, 95% of the time you get good interval estimates that capture the population proportion". From a relative frequency perspective, the first sentence in Hypothetical Student 2's response means the same thing as the second sentence. This is a powerful understanding of confidence level that I argue should be developed in students.

Incoherent Hypothetical Student Interpretations

Liu (2004) also suggests several alternative interpretations of confidence interval and margin of error. Of course, she includes among her list of different interpretations the classic interpretation of confidence intervals that does not entail an image of repeated sampling; that is, the view that the true population parameter *is* inside the interval, *sample statistic* ± *margin of error*. Hypothetical Student A's interpretation, "The margin of error being 4% means that between 32% and 40% of all Oregon voters believe an income tax is necessary" exemplifies this view of margin of error and confidence intervals. I consider this view problematic from an educational standpoint because it: (a) does not explicitly mention the level of confidence; and, (b) contains an implicit assumption that the population parameter moves. From a statistics education stand point, it is preferable for students to develop a conception that supports the interpretations of both Student B and E because taken together such a construal provides students the understanding of the power of sampling distributions in making statistical inferences. Also, an understanding of Student B's and E's interpretations

entails, at least implicitly, an expectation of variability among sample statistic values. I added Hypothetical Student 1's interpretation of confidence level as an analog to Student A's interpretation of margin of error. Hypothetical Student 1's response is a common interpretation for confidence level, in the same way that Hypothetical Student A's interpretation of margin of error represents a common view. Thus, Hypothetical Student 1's interpretation is also troubling because the confidence is placed within the particular interval calculated from the sample.

In addition, Liu (2004) provides other alternative interpretations of margin of error and confidence intervals that include a conception of repeated sampling, but represent incoherent ways of thinking about confidence intervals. For example, Liu suggests one could interpret margin of error with an image of the repeatability of the sampling process, but believe that the interval obtained in a particular sample will contain some percentage, x, of all the other sample statistics; that is, *sample statistic $\pm$ margin of error* captures *x% of all sample statistics*. I created Hypothetical Student D's interpretation of the poll's margin of error, "if you performed repeated samples of 500 voters, the proportion of voters in favor of sales tax in these samples would fall within the interval 32% to 40%, the majority of the time", as a reflection of this incoherent concept of margin of error. I added two additional incoherent images of margin of error and confidence intervals. First, Hypothetical Student F's interpretation of the poll's margin of error, "the interval 36% $\pm$ 4% has a high probability (approximately 95%) of being repeated if the sample was repeated", suggests an image that entails a view of repeated sampling, but expects the particular interval computed from the one

223

particular sample to have a high probability of being repeated. Second, in Hypothetical Student C's interpretation of the poll's margin of error, "the interval 32% to 40% will be off about 4% of the time, or 4 out of 100 times", it is not clear whether an image of repeated sampling is present. Hypothetical Student C also uses the 4% margin of error in two different ways: Student C correctly used the 4% margin of error to compute the upper and lower limits of the confidence interval, (32, 40), but then incorrectly used the 4% to suggest a level of confidence in the chosen interval's predictive power ("off 4% of the time").

5.3.3 TAs' Reasoning about the Gallup Poll Task

In this section I examine TAs' responses to the Gallup Poll Task and to the different hypothetical student interpretations. I suggest that TAs appeared to reason on a continuum from strong images of repeated sampling in relation to confidence intervals to no image of repeated sampling. I begin this section with a discussion of TAs' initial responses to the Gallup Poll Task, followed by a discussion of how TAs interacted with the hypothetical student interpretations. This order allows me to provide a picture of how TAs' interpretations changed as a result of their thinking about the hypothetical student interpretations. Also, this presentation provides me the opportunity to illuminate the continuum of TAs' responses, ranging from strong images to no images of repeated sampling. Please note that I do not discuss each TA's response to every hypothetical student, as this would be too time consuming. Rather, I

provide responses that highlight key aspects of TAs' thinking related to this concept of repeated sampling.

TAs' Initial Responses to the Gallup Poll Task

Amanda is the only TA who began her discussion of the Gallup Poll Task with an interpretation of margin of error and confidence that provides some indication that she held an image of repeated sampling.

Interview 3: Gallup Poll Task:

> **Amanda:** Well, in the most simplistic terms that means that 36% is our point estimate and our cushion provides room for 32 to 40%. It doesn't tell us much about what level of confidence they're using…. There's error involved in the sampling process, it's not an exact representation of your population. So with whatever level of confidence they chose we're looking at an estimate of the proportion being between .32 and .40.

Amanda's unprompted initial discussion of margin of error raises the notion of confidence and of "error involved in the sampling process". Although this information is not sufficient for inferring that Amanda holds an image of repeated sampling, she is the only TA to raise the issue of confidence in relation to the *sampling process* rather than in regards to the *specific interval* found from the sample statistic in her initial utterances.

In comparing Amanda's initial response to the Gallup Poll Task with Andy's and Joe's responses, there appears to be a subtle, yet significant, distinction in their interpretations. Andy's initial interpretation of margin of error in the Gallup Poll Task is provided in the following exchange.

Interview 3: Gallup Poll Task:

225

> **Andy:** Okay. So there's some true value. And so this is establishing this confidence interval so we're saying our confidence interval is 36 minus 4, which would be 32%, to 36 plus 4, or 40%. So we are saying with some degree of unstated confidence, nobody's saying what it is at this point, the true value is between 32 and 40.

Joe's initial response to the Gallup Poll Task was similar to Andy's in that he came up with the interval around $\hat{p} = 36\%$ and indicated that there is some uncertainty in this interval.

> Interview 3: Gallup Poll Task:

> **Joe:** We're between 32 and 40 kind of in a statistical way. We can never be entirely confident on what the situation is when we are not reading all the, everyone's opinion in the state, but we're fairly confident in our confidence interval that it's about 36%.

Andy's and Joe's interpretations of margin of error suggest that they are aware that the poll has some unstated level of confidence and that the interval (32, 40) may not contain the true percentage of Oregon voters that believe a state sales tax is necessary. However, it is not clear from either excerpt whether Andy or Joe held implicit images of repeated sampling. In fact, Andy's utterance, "with some level of confidence the true value is between 32 and 40", suggests an image that the particular interval obtained contains the population parameter. However, Andy and Joe used language for describing confidence intervals that is commonly employed by those with a frequency interpretation and those with a subjective interpretation of probability; thus, it is difficult to tell how Andy and Joe actually conceptualize the situation. The difference between Amanda's initial response and the initial responses given by Andy and Joe is subtle. Yet, Amanda's reference to error in the *sampling process*, compared to Andy's

and Joe's references to error in the *particular interval* calculated, *could* be an indication that Amanda has different mental schemas around the concept of confidence interval than do Andy and Joe.

Sandy's and Sam's initial interpretations are distinguishable from Amanda's, and Andy's and Joe's interpretations because Sandy and Sam did not make an explicit reference to the confidence level, nor to the concept of repeated sampling. Sandy's initial interpretation of margin of error and confidence interval are shown in the following exchange.

Interview 3: Gallup Poll Task:

Sandy: …So 36%, so I know $\hat{p}$. Let me try, if I will write the confidence interval for proportion [***See Sandy's work, Figure 40***]. I'm not very sure I don't remember if this is it. [***Long pause***]. Confidence interval for proportion. I know it's a *z*. Ahh, it's $\hat{p}$, $\hat{q}$, over *n* [***instead of p and q which she had before***].

Figure 40: Sandy's Confidence Interval
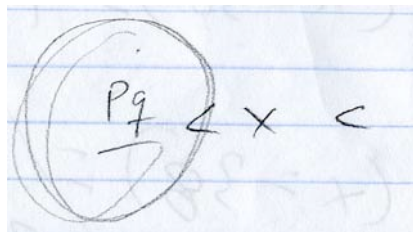
Sandy's response continued: …So margin of error, we call this margin of error [***underlines the*** $z\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$]…. Okay, okay, okay Jennifer. So 36% plus/minus .04 and I get the confidence interval. So I have, when I do the minus, .32, when I do the plus, .4. So what happened was, the margin of error helped me to get the confidence interval because I know the $\hat{p}$. So I get the confidence interval, so now I know that the proportion of Oregon people who think a sales tax is necessary to overcome the budget problems is between 32% and 40%.

Sandy approached the problem by trying to recollect the formula for confidence interval for proportion. After Sandy recalled the formula, she provided a common interpretation for confidence interval, which suggested the population parameter *would be* in the interval 32 to 40, as seen in the last highlighted utterance from the previous excerpt. Sandy made no mention of how confident she could be in her prediction or any other additional utterances that would be helpful for understanding whether or not she had an image of repeated sampling.

Like Sandy, Sam attempted to use a formula for confidence intervals in order to compute the confidence interval for the Gallup Poll Task. Sam tried to describe the process of adding and subtracting 4% to $\hat{p} = 36\%$, unfortunately he could not recollect the exact formula for a confidence interval for proportions, and he could not provide me with the interval (32, 40). It seemed that Sam did not have an alternative way of articulating his thinking about confidence intervals, other than through a formula. Thus, from his initial utterances alone, I was unable to gather enough information to speculate about his understanding of confidence intervals. Figure 41 shows the progress Sam made on re-creating the formula for confidence intervals for proportions.

Figure 41: Sam's Confidence Interval

During our conversation Sam displayed a highly procedural understanding of confidence intervals and suggested he needed to know the formula for confidence intervals involving proportions in order to discuss the Gallup Poll Task. Perhaps Sam does not view confidence intervals as being of the form, *sample statistic* ± *margin of error*, because Sam appeared to focus on the specific formula for a confidence interval for proportions.

TAs' Responses to Hypothetical Student A's Interpretation of Margin of Error

As Amanda interacted with the hypothetical student responses, this image of repeated sampling became more explicit in her utterances. Take, for example, Amanda's reaction to Hypothetical Student A's interpretation, that the margin of error being 4% means that between 32% and 40% of all Oregon voters believe an income tax is necessary.

Interview 3: Gallup Poll Task:

> **Amanda:** I think that, I think that this is a very common interpretation. I don't think it's fully accurate, but I think conceptually it's how we process information, like 32% plus or minus 4%.
> **Interviewer:** What do you think is inaccurate about it?
> **Amanda:** That the true proportion might not be between 32 and 40%.
> **Interviewer:** And that's not explicit in the statement?
> **Amanda:** Right. That with a certain level of confidence the true proportion will be between 32 and 40%, but the truth of the matter is that the true proportion is either in this interval or it's not and this way of talking about it I think is very common…. I think it's a very nuanced idea to start talking about the fact, and you have to, but still it's very nuanced to discuss the fact that what does a confidence interval really mean and it takes a while…. You know let's assume 95% confidence, with 95% confidence we can assume the true proportion would be between 32 and 40%, but that we're actually talking about 95% of all samples would capture the true proportion and either our sample did or it did not…. I think there is nuance

to it that leaves you with the impression that the true value is definitely between 32 and 40%, and it might be but with a certain level of confidence it will be.

In this excerpt Amanda indicated that Hypothetical Student A's interpretation is incomplete because it does not mention confidence level or the fact that our interval may not contain the true proportion. In addition, the highlighted utterance at the end of the exchange is where Amanda provided an explicit image of repeated sampling, indicating that 95% of all samples would capture the population parameter.

Unfortunately, Hypothetical Student A's response was not as useful in providing deeper insight into the mental schemes of Andy, Joe, Sandy, and Sam. For example, Andy and Joe both commented that Hypothetical Student A neglected to mention confidence level, which they had previously addressed in their initial interpretations, but there was nothing in Hypothetical Student A's response that pushed Andy or Joe to make explicit their notions of the process of creating confidence intervals. Andy's and Joe's responses to Hypothetical Student A's interpretation are shown in the following exchanges.

Interview 3: Gallup Poll Task:

Andy: The margin of error being 4% means that between 32 and 40% of all Oregon voters believe an income tax is necessary. Yeah they've got part of it. They talked about that's the interval. They didn't talk about confidence nature of it.

Interview 3: Gallup Poll Task:

Joe: Yeah, if this question was worth 10 points I would say this is 8 or 9 points.
Interviewer: Okay. How come?

Joe: Because between makes me think that they haven't quite got it. The ±4% doesn't say it's necessarily between here [***the 32 to 40% interval***].…But the actual number of Oregon voters who believe that, is *likely* [***Joe's emphasis in phrasing***] between 32 and 40%. And the way they've stated this I, I haven't seen that *likeliness* captured and that's an important part of it.

Interviewer: Okay, so it's like they are losing a few points in your mind because they're saying it with *surety* and not *likelihood*?

Joe: Yeah…. It's still possible you know that we've just polled people in downtown Portland and they all live in a commune, and it doesn't really reflect the entirety of, of voters in Oregon.

Andy and Joe noted that Student A failed to mention how confident he/she was that the interval calculated would capture the population parameter. Yet, Andy's and Joe's initial utterances and their responses to Student A's interpretation of margin of error remain ambiguous on the issue of repeated sampling. Taking a conservative approach in characterizing their understandings of confidence intervals suggests that their confidence is in relation to the *specific* interval (32, 40) gathered from this one sample, rather than confidence in the sampling process.

A second point worth noting in Joe's excerpt is his response to why a confidence interval may not capture the population parameter. In the final highlighted utterance, Joe suggested biased polling as a reason for why the interval might not be a good reflection of the population parameter. This utterance leads me to wonder how Joe understands the type of error represented in a particular confidence level. That is, does Joe view such error as pertaining to sampling bias or to natural random error? I am not addressing this question in my study, and although Joe's utterance provides some evidence that he views the 5% error (in a 95% confidence interval) to be a result of sampling bias, it is unclear whether he also thinks about natural random error.

Hypothetical Student A's response was also not useful for gaining a deeper understanding of how Sandy and Sam conceptualized confidence intervals. In essence, Hypothetical Student A's interpretation corresponded with Sandy's initial interpretation. Although Sam had not provided a clear articulation of how he interpreted confidence intervals, I suspected Sam had a similar notion of confidence interval as Hypothetical Student A because of his focus on finding a formula for computing the confidence interval. Given that Hypothetical Student A's interpretation was similar to Sandy's and Sam's initial interpretations, the conversation did not lead toward an explicit conversation around their interpretations of confidence level. At that point during the interview, all that could be inferred from my conversations with Sandy and Sam was that their interpretations of confidence interval most closely matched Hypothetical Student A's interpretation. Take for instance the following exchanges, which demonstrate that Sandy and Sam found Student A's interpretation acceptable because it was essentially the same interpretation they were trying to articulate to me in their initial response to the task.

Interview 3: Gallup Poll Task:

> **Interviewer:** Student A says that the margin of error being 4% means that between 32% and 40% of all Oregon voters believe an income tax is necessary.
> **Sandy:** Yeah because, so he pinpointed correctly the confidence interval. And he specified all Oregon voters.
> **Interviewer:** Meaning?
> **Sandy:** The true population proportion.

Interview 3: Gallup Poll Task:

Sam: Oh, Okay. Yeah. Yeah, that was what I was explaining the first time. …Yeah, I think that should be this confidence interval. You know you asked me about confidence interval before and I wasn't able to explain. So that should be the interval, confidence interval goes with margin of error. So you are bringing my mind back to this.

Based on Sandy's and Sam's initial interpretations of the Gallup Poll Task and their acceptance of Hypothetical Student A's interpretation, it appears that Sandy and Sam do not think about confidence level, nor do they appear to have mental schemes that entail repeated sampling. Rather, Sandy and Sam appear to believe that the population parameter *is* contained in the interval obtained from the sample statistic 36%. Sandy's and Sam's utterances so far indicate a perspective more closely aligned with a subjective interpretation of probability. However, it is wise to be cautious with such an interpretation of these TAs' thinking because Student A's interpretation is often used as a shorthand interpretation by statisticians, who, in fact, hold a long-term relative frequency perspective.

TAs' Responses to the other Hypothetical Student Interpretations

The remaining Hypothetical Student interpretations, Students B-F, entailed some aspect of repeated sampling. These tasks pushed TAs toward being more explicit about their thinking around the concept of confidence level and enabled me to gain deeper insight into how they conceived of the role of repeated sampling in the creation of confidence intervals. For example, Sandy's, Andy's, Sam's and Joe's responses to Hypothetical Student B's interpretation helped to clarify where their reasoning would be placed on a spectrum ranging from strong images to no image of repeated

sampling. Also, TAs' responses to Hypothetical Student E and Hypothetical Student's

1 and 2 provided a more detailed picture of TAs' reasoning about confidence intervals.

I begin with a discussion of how Sandy, Andy, Sam and Joe responded to

Hypothetical Student B's interpretation. I follow with a discussion about TAs'

responses to Hypothetical Student E and end with a discussion about TAs' responses

to Hypothetical Students 1 and 2.

TAs' Responses to Hypothetical Student B's Interpretation

As Sandy investigated Hypothetical Student B's response, there was increasing

evidence that her concept image of confidence intervals *did* include confidence level

and repeated sampling, although these ideas were not explicit in her initial utterances.

As Sandy continued to engage in the Gallup Poll Task, and the issues of confidence

level and repeated sampling were addressed in the hypothetical student responses, she

began to make these ideas explicit in her utterances.

When I asked Sandy about Student B's interpretation of margin of error, and

confidence intervals she initially appeared perplexed. The following exchange shows

her initial reaction was to reiterate her own interpretation of confidence interval.

Interview 3: Gallup Poll Task:

Interviewer: So, Student B says: we don't know if the interval 32% to
40% contains the true percentage of voters that believe an income tax is
necessary, but if we sample 100 times, about 94 of those times the interval
would capture the true percentage of voters.
Sandy: Oh my God I am tired and this is so wordy I don't know.
[*Laughs*]. [*Sandy reads the problem out loud again*] We don't know if the
interval contains the true percentage of voters? But we know that this
should contain the true percentage of voters because [*pause*].

234

> **Interviewer:** Why, what tells us we know that?
> **Sandy:** Okay. So you have a confidence level. You are 95% confident that the true proportion is there [**Sandy points to the CI**]. Of course there is this 5% where you can be off. So I'm not sure right now if "we don't know" refers to that [**the 5%**]….When you speak about confidence interval you have to state how confident you are that that is where your true population parameter is located.

Prior to investigating Student B's interpretation of margin of error, Sandy did not

mention the issue of confidence level. It appeared that she either believed the interval

obtained *would* contain μ, or she implicitly had some unstated level of confidence in

mind but did not articulate it. However, as Sandy considered Student B's

interpretation, the issue of confidence level surfaced explicitly. In the first highlighted

utterance she suggested that the interval *should* contain the true percentage of voters,

but she paused and then explicitly mentioned that this would be true with some level

of confidence. The third and fourth highlighted utterances give some indication of how

Sandy made sense of confidence level. Sandy stated that the 95% confidence interval

means that you are confident that the true proportion *is* inside the interval. Again, this

interpretation makes it difficult to determine whether or not Sandy's image of

confidence level entails repeated sampling.

As Sandy continued to examine Student B's interpretation, she began to express a

more explicit image of confidence level that *did* entail repeated sampling.

> Interview 3: Gallup Poll Task:

> **Interviewer:** So, is there a way for us to figure out what the confidence level for this problem would be?
> **Sandy:** Yes, yes we could. Okay so we know this number .04. We know that [**points to $\hat{p}$**], we know that [**points to $\hat{q}$**], we know this [**points to n**]

we can find z [*looking at her expression for margin of error in the CI*].
And if we find z, we can get alpha.
Interviewer: So then when they say 94 times out of 100 could they mean
something like your 95% that you were using when you talked about this?
Sandy: But if we sample 100 times, but I'm not sure about 94. Ahh, I
believe that maybe that is what he is trying to say. If I sample 100 times,
94% of the times the interval will capture the true percentage. Ahh, I
believe that's it. Maybe if I will find this z and I will see that 94% is the
confidence level that means that is what he wanted to say. That you cannot
be sure 100 percent, but in 94% of the cases you could be.
Interviewer: Okay. Is that what 95% means when you were putting it there
[*referring to her earlier comments on 95% Confidence level*]?
Sandy: Yeah. It's 95% of the time, so you know you want to repeat and
repeat and repeat.
Interviewer: The sampling process?
Sandy: The sampling process. Compute this $\hat{p}$, compute the confidence
interval, compute the confidence interval, compute the confidence interval
and you will see that if you use a certain level you will see that in 95% of
the cases that's, you will know in 95% of your cases that's where the true
percentage of the population will be.

In this last exchange, Sandy makes specific reference to 95% confidence as being

related to the *sampling process*, but prior to this excerpt her responses suggested that

the interval obtained in the *particular* sample would contain (with some level of

confidence) the population parameter. Sandy's last highlighted utterance provides

strong imagery of repeating the sample 100 times, and that 95 of those 100 times the

interval estimates will capture the population parameter. It is difficult to say if her shift

to repeated sampling occurred as a result of my questioning or if my questions

prompted her to make her tacit assumptions more explicit. In any event, this excerpt

provides explicit evidence that Sandy was thinking about repeating the sampling

process and that her understanding of confidence level was in regard to that sampling

process.

Hypothetical Student B's interpretation of margin of error also provided more explicit evidence that Andy, Joe and Sam *did not* have concept images of repeated sampling in relation to confidence intervals. For example, as Andy interacted with the different hypothetical student responses he appeared to wonder why each of these students discussed the idea of repeated sampling, and he wondered where students would have developed such an image. His reactions provided stronger evidence suggesting that he imagined the confidence level related to the *specific* interval gathered from a single sample. The following exchange is Andy's reaction to Hypothetical Student B's interpretation of margin of error.

Interview 3: Gallup Poll Task:

Interviewer: Okay, another student said we don't know if the interval 32% to 40% contains the true percentage of voters that believe an income tax is necessary, but if we sample 100 times about 94% of those times the interval would capture the true percentage of voters.

Andy: Wait a second. [***Long pause***]. Yes, it's true that we don't know that this interval contains the true percentage. If we are going to sample 100 times, 94% of those will be in this interval. [***Andy reads the second part of the student's response aloud***]. Yeah, that's not what it says though. [***Long pause***].

Interviewer: That's not what confidence interval says or…

Andy: Yeah that's not what this confidence interval says. It doesn't say a thing about re-sampling. It doesn't imply re-sampling. It talks about the margin of error and I'm going to stick with my definition [***laughs***] that it's not related to the re-sampling or the hypothetical re-sampling of it.

In this exchange, Andy appeared surprised to see an interpretation of confidence interval based on repeated sampling (hypothetical or otherwise) and he believed that this student's response was incoherent.

I continued to question Andy on the idea of repeated sampling in my follow-up questions in order to establish if he had any conception that repeated sampling is an implicit part of the theory behind the concept of confidence intervals. The following conversation took place.

Interview 3: Gallup Poll Task:

> **Interviewer:** So I guess my question is, you are saying in here [*in the Gallup Poll statement*] that there is sort of an implied confidence level. Could there also be an implied idea of repeated sampling?
> **Andy:** There could be, but I've never understood it to be that way. Maybe it is and I've always just misinterpreted it.
> **Interviewer:** Okay, but at the moment you're kind of thinking…
> **Andy:** At the moment I need to berate them for being totally wrong [*laughs*].
> **Interviewer:** [*Laughs*]. Okay, because they are bringing in this idea of repeated sampling?
> **Andy:** Right.
> **Interviewer:** So you're saying sure they're right here we don't know if the 32 to 40 is going to capture the true proportion?
> **Andy:** Right we don't, we never know. That's the whole point about confidence – we're pretty sure, but we don't know.

In this exchange, Andy explicitly stated that he did not think about the idea of repeated sampling in relation to confidence intervals. In the last highlighted utterance, Andy's remark, that the interval obtained from the sample might not capture the population parameter but that the confidence level indicates how likely it is, suggests that his confidence is in the *particular* interval computed from the sample.

Like Andy, Sam appeared perplexed by the notion of repeated sampling in relation to confidence intervals.

Interview 3: Gallup Poll Task:

Interviewer: So Student B said that, 'we don't know if the interval 32% to 40% contains the true percentage of voters that believe an income tax is necessary, but if we were to sample 100 times, about 94 of those times this interval 32% to 40% would capture the true percentage of voters'.
Sam: Okay, so why should you go and sample 100 times? I mean that would be a waste of time.
Interviewer: It would be a waste of time? So is what they are saying incorrect or just that it's not feasible?
Sam: …My view, I would use the formulas, like if you know margin of error. I mean I would go for using the margin of error for calculating this.
Interviewer: Okay. So between Student A's response and Student B's, you would say Student A is a better response?
Sam: Yeah, Student A was the response I gave as one of my answers. At first I wasn't that sure, but definitely I would go with that.

Sam did not like Student B's interpretation because he thought that sampling 100 times would be a waste of time. Sam's response to Student B's interpretation seems to suggest that he did not have an image of repeated sampling in this context. Yet, it could also indicate that he believed that Student B was literally suggesting sampling 100 times, rather than a hypothetical image of repeated sampling that supports the conception of confidence level. It is also interesting that Sam indicated he would stick with the formula for calculating margin of error. Sam's initial attempts to recollect a specific formula for confidence intervals for proportions, followed by his response that it was best to stick with the formula, indicates that he was not comfortable providing interpretations for what these interval estimates represent. Rather, he was more concerned with using a procedure to find an interval estimate.

Joe's response to Hypothetical Student B's interpretation suggests that he was perhaps familiar with the notion of repeated sampling in relation to confidence intervals, but that he did not think it was a constructive way of interpreting confidence

intervals. In the following exchange, Joe's utterances suggest that his mental scheme

fit more closely Student A's, with the added caveat that one can never know for sure,

but can be pretty sure the interval we get around our sample statistic captures the

population parameter.

Interview 3: Gallup Poll Task:

> **Joe:** …This [**Student B's response**] is a half way formal answer, if that makes sense. That's how I think of it. This [**Student A**] I think is, in a certain sense, a better answer because it captures the meaning of what's going on here.
> **Interviewer:** Student A does?
> **Joe:** Yeah. And like the realistic meaning when I read a newspaper and I'm thinking about what statistics reported mean. Yeah, because if I'm looking at a poll and I'm sampling people from the poll, …I'm not thinking what happened if I took another sample of 100 voters. That is important from a statistics and a test taking sense, but in a reading a newspaper and figuring out the world sense that's not the useful way to think about it.
> **Interviewer:** Why?
> **Joe:** The useful way to think about it is how many people think a sales tax is necessary. What does this statistic mean? 36% think it is, with a margin of error of ±4%.
> **Interviewer:** So there is no image of sampling, the sample being repeated in this way. It sort of sounds like you're saying you're trying to make a decision – do Oregon voters want a sales tax based off this one sample?
> **Joe:** Right….This [**Student A's response**] captures the meaning and the interpretation that's important contextually for an educated layperson.

What is particularly compelling in this exchange is Joe's argument that Student

A's response captures the meaning of confidence interval that is important for students

and/or educated citizens. What seems to be important for Joe when reading a poll is to

be able to predict, with some level of confidence, how likely the population parameter

is to be contained within the interval estimate. This last excerpt suggests that Joe's

mental schema appears to be more closely aligned with Student A's interpretation,

240

which is a subjective interpretation of probability. From both a pedagogical and statistical perspective, I disagree with Joe's argument that Student A's interpretation captures the meaning that is important for an educated citizen because I believe that students should have a fundamental understanding of the conceptual process entailed in sampling and statistical inference. This point is discussed further in Chapter 6.

TAs' Responses to Hypothetical Student E's Interpretation

Hypothetical Student E's interpretation of margin of error provided another opportunity to examine TAs' understandings of confidence intervals. For example, Amanda's frequency interpretation of confidence level became more apparent in her discussion of Hypothetical Student E's interpretation of the Gallup Poll Task. Recall that Student E's interpretation is an alternative coherent image of confidence interval, in which the visualization is on the distribution of sample statistics. From my discussions with the TAs, Amanda was the only one who appeared to connect Student E's interpretation to Student B's, and to suggest that Student E's was coherent and consistent with Student B's (the more standard interpretation of confidence interval found in introductory texts). The following exchange shows how Amanda used the concept of sampling distributions in order to make her decision regarding the validity of Student E's interpretation.

Interview 3: Gallup Poll Task:

Amanda: I'm trying to visualize. I've now drawn a picture of a distribution of $\bar{x}'s$ centered at μ….And if I mark that margin of error around μ then absolutely what I've just captured is the 95% of $\bar{x}$ values surrounding μ because I've counted the appropriate number of standard

deviations on either side of μ away to capture 95%. So that actually would be 1.96 standard deviations around each side of μ and that's exactly what that would do is capture 95% of the $\bar{x}$…. And if I bring it back to μ because I'm more comfortable that way, then this [**pointing to the graph she just sketched, see Figure 42**] would be the sampling distribution of $\bar{x}$.

Figure 42: Gallup Poll Task – Amanda's Sampling Distribution
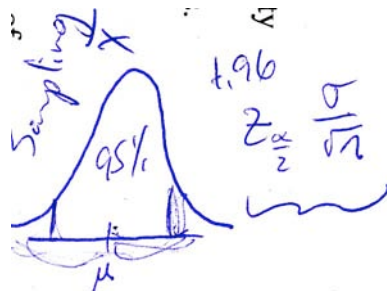
Amanda's Response Continued: So sampling for $\bar{x}$ centered at μ and the margin of error is counting so many standard deviations away from, well whatever you decided to center it around. In the confidence interval it's counting standard deviations centered at $\bar{x}$. But if you think about the actual sampling distribution, and then counting out the margin of error on either side, then precisely what you've done is captured 95% of the most common values. This is the link when you're talking about μ's. It's the link between two-tailed test and confidence intervals. So I feel comfortable with him saying then that 95% of my sample statistics, so my point estimates, would fall into this region surrounding my true parameter.

In the first highlighted portion of the excerpt, Amanda suggests that she needed to visualize what this student is saying. She drew a sampling distribution of sample means, shown in Figure 42. In the second highlighted portion of the excerpt, Amanda explicitly made a connection between the sampling distribution and a 95% confidence interval. That is, she was able to display the distribution of sample statistics and examine the percentage of those statistics that should fall within a certain distance of the population parameter. This is a strong indication that Amanda's mental scheme of

242

confidence intervals entailed sampling distributions and her notion of confidence level was with reference to the process by which intervals are produced.

Sandy approached Hypothetical Student E's interpretation differently than Amanda. Sandy was able to show mathematically that Student E's interpretation was logically equivalent to the standard textbook interpretation. Yet, Sandy ultimately rejected it as an acceptable interpretation because Student E's interpretation was not how textbooks interpreted margin of error. In addition, Amanda reasoned through Student E's interpretation from the perspective of the sampling distribution whereas Sandy, being proficient with symbolic manipulation, reasoned about it from an algebraic point of view, shown in the following exchange.

Interview 3: Gallup Poll Task:

> **Sandy:** Yeah. Can I say vice a versa? Can I move this here? [**_Long pause_**]. So I will get $\hat{p}$. Okay, I will just take it by each part…. Hmm, yeah you can say that. [**_See Figure 43 for Sandy's written work_**].

Figure 43: Gallup Poll Task – Sandy's Work

243

Although Sandy concludes that Student E's statement is equivalent to Student B's interpretation, she does not appear comfortable concluding that Student E's interpretation is statistically sound.

Interview 3: Gallup Poll Task:

> **Sandy:** Yeah, but normally we want to say that the true one is between the statistic ± the error. Because your interest is, so what I'm trying to say to my student is that you're not interested in what's up with the statistics for the sample because you know what is going on. If you want, go collect a sample and you know exactly the value of $\bar{x}$, σ, whatever. Your interest is about the population parameter – that's what you don't know. So I believe that he just reversed here. 95%, first of all it's not 95, its 94, and that sure, that the population proportion will fall within ± 4 of the sample p....Yeah, they wanted to say something here.... The sample statistics is between p +E and p-E, but we don't interpret like that because we don't know p and you know $\hat{p}$. So I don't agree with this Jennifer.

Sandy indicated that she would not accept this interpretation because she believes the student is confused about the parameter of interest. The first highlighted utterance from this excerpt suggests that Sandy expects to see the form typically displayed in a statistics text because the primary interest is in the population parameter. Sandy's utterances in this last exchange suggest that because we do not know the population parameter, it does not make sense to talk about placing an interval around the unknown population parameter. This suggests that perhaps Sandy is not connecting the confidence interval to the sampling distribution. Sandy determined that Student B's and E's interpretations were equivalent algebraically, whereas Amanda determined the equivalence through the use of the sampling distribution of sample statistics; perhaps Sandy's image of the connections between sampling distributions and confidence intervals is not as robust as Amanda's image in this context.

Andy, Joe and Sam did not think that Student E's interpretation was a valid or coherent interpretation. For Joe and Sam, Student E's interpretation deviated from Student A's interpretation, and they both identified most closely with Student A's interpretation. In addition, Andy and Sam vehemently opposed any interpretation that mentioned repeated sampling. Andy did mention that Student E's interpretation was a nice, natural way to think about margin of error, but he thought it was an incorrect way of thinking about it.

Interview 3: Gallup Poll Task:

> **Andy:** …Okay, so they're assuming it's 95, I think. Fall within ± 4%, now you see ± 4% of the unknown population parameter that's got things a little confused because this is 36. It's the 36 that's ± 4….The unknown thing is just that, it's unknown. It's not as though we are going to get within ± 4% of it. We're here [*pointing to interval around 36, the sample proportion*]…. And this [*Student E's statement*] seems like a nice way to talk about it. It's like well we are chasing this thing and we can get within ± 4% of it, right, with 95% confidence, right. And it's a good way to think about it. It's just not how it's defined….It's easy to hold in your head. I'm chasing my goal. I'm going to get within this percentage of my goal, great. Well, bad luck because that's not how it's defined.

Andy suggested that Student E's interpretation is a natural way of thinking about confidence intervals because we want to be able to say something about the population parameter, but he did not believe that this interpretation was coherent. There is some evidence in this excerpt that suggests that Andy did not have an image of repeated sampling and the distribution of sample statistics in relation to the concept of confidence intervals, which could account for why he did not find Student E's interpretation coherent. For example, the second and third highlighted utterances from

the excerpt above do not suggest the imagery of a collection of sample statistics within a certain distance of the population parameter. Rather, the imagery in those highlighted utterances is suggestive of a single sample statistic from the particular sample within four units of the unknown population parameter. In particular, Andy's utterance "we are chasing this thing and we can get to within $\pm 4\%$ of it, right, with 95% confidence, right" is suggestive of the latter imagery.

TAs' Responses to Hypothetical Students 1 & 2

Although I intended the use of Hypothetical Student 1's response for confidence level to be inconsistent to that of Hypothetical Student 2's response, Amanda did not see it that way. To me, Hypothetical Student 1's response is suggestive of confidence in the *particular* interval obtained in the sample and Hypothetical Student 2's response is suggestive of confidence based in how many of these intervals capture the population parameter over the long run. Yet, the following excerpts show that Amanda perceived the two interpretations as the same.

Interview 3: Gallup Poll Task – Response to Student 1:

Amanda: This is okay. If 95% of the point estimates are going to fall within the margin of error of the true parameter, then 95% of the intervals are going to capture the true parameter.

Interview 3: Gallup Poll – Response to Student 2:

Amanda: …But then this explanation that 95% of the time you get good interval estimates that capture the population proportion. I feel like that's saying the same thing the other student said.

In justifying her approval of Student 1's response, Amanda appeared to go back to the duality expressed between Student B's and E's phrasing of margin of error to make a case that Student 1's interpretation of confidence level was consistent. Amanda stated that 95% of sample statistics fall within a certain distance of the population parameter, and that means the same as 95% of the confidence intervals capturing the population parameter. Thus, it appears that Amanda read Student 1's interpretation as saying that 95% of the confidence intervals capture the population parameter, not that there is a 95% chance that the population parameter, p, is inside the *particular* confidence interval obtained from the sample.

Like Amanda, Sandy thought that Student 1's interpretation was acceptable; however, she found it incomplete. In the next exchange, Sandy's image of repeated sampling is quite explicit.

Interview 3: Gallup Poll:

> **Sandy:** Yeah, and that's how we interpreted the confidence interval, you know in my class [*a statistics class Sandy was enrolled in*]. Because the more textbook interpretation is like in 95, if you repeat the sampling process 95% of the times, you'll capture in your confidence interval the true parameter, population parameter. We just had this discussion in the stats class I'm taking, and I learned of the repeating the sampling process, but our professor said that it is okay to use this, 'I am 95% confident that' [*points to the student's interpretation of confidence level that we just read*].
> **Interviewer:** That the interval you just got captures the population proportion?
> **Sandy:** Yeah, yeah. And I believe that maybe this is not very exact.
> **Interviewer:** So why, what is lacking in this?
> **Sandy:** I am 95% confident, because you miss saying that you have to sample many, many times. You know.

Sandy indicated that Student 1's interpretation was acceptable and was even used as a convenient shorthand interpretation by her professor in a statistics course. This convenient shorthand interpretation is one reason why it is so difficult to tell from TAs utterances whether or not the image of repeated sampling is present. The vocabulary of the shorthand interpretation does not make explicit the assumptions of repeated sampling, and Sandy recognized that this shorthand interpretation neglects the concept of repeated sampling. Sandy indicated that maybe such a discussion already took place in the classroom and that perhaps the student was using this abbreviated version, but she also suggested that a student might miss this interpretation.

Of the five TAs I interviewed, Andy appeared to agree most closely with Hypothetical Student 1's interpretation of confidence level specifically because the interpretation did not make mention of repeated sampling.

Interview 3: Gallup Poll Task:

> **Andy:** So a student says that a 95% confidence level means that you can be 95% confident that the particular interval found in the survey, that one [*points to the interval from the Gallup poll*], captures the population proportion. Well I agree, but only because it's almost a truism. You know it's like yes that's absolutely true that blue equals blue, but I have something to tell you about this like you're not really saying anything new…. But at least they've got this other part. They've got that the population proportion is in that interval and the other student's [*referring to Students A-E on previous page with Gallup poll*] don't have that concept.

Andy found this student's statement to be redundant, but that it was better than any of the prior student interpretations. In the second highlighted utterance Andy suggested that Student 1 "gets" something that the other students did not "get". What Student 1

248

understands, from Andy's point of view, is that the population proportion *is* in the interval (with some amount of confidence). I believe that overall, Andy's response to this task strongly suggests that he did not have an image of a distribution of sample statistics in relation to confidence intervals.

Hypothetical Student 2's interpretation was given to TAs because it expresses an interpretation of confidence level in relation to confidence in the sampling process, rather than the particular interval captured from the sample. Yet, Amanda was the only TA to recognize this phrasing. In addition, she recognized that the choice of wording could be construed by others as questioning the survey process, rather than the sampling process. The following exchange shows Amanda's response to Hypothetical Student 2's interpretation.

Interview 3: Gallup Poll – Response to Student 2:

Amanda: [***Long pause***]. Yes, I think the phrasing 95% confident in the estimation process is a little weird.
Interviewer: What do you find weird about that one?
Amanda: The estimation process? That I'm declaring some amount of confidence in what the way the sample was gathered? [***Laughs***] Yeah, that speaks to me on levels that aren't really about crunching the numbers, but I don't know almost the integrity of the people who performed the survey or something [*laughs*]. But then this explanation that 95% of the time you get good interval estimates that capture the population proportion. I feel like that's saying the same thing the other student said.

Given that Amanda read into Student 1's interpretation a tacit assumption of repeated sampling, she perceived it to be essentially the same interpretation as Student 2 offered and she saw no inconsistencies, as evidenced by her utterances in relation to

Student 1 (at the beginning of this subsection) and the last two sentences in the previous excerpt.

However, Amanda argued that Student 2's interpretation could be construed as questioning the polling procedures. She suggested that the phrasing "confidence in the estimation process" made her think that the student was questioning the integrity of the people conducting the poll. Although Amanda was able to see past this source of confusion based on the rest of Student 2's interpretation, this phrasing appeared to be bothersome to the other TAs as well. For example, Andy, Joe, and Sandy were confused by Student 2's response and suggested that the "process" of the survey is not under question when interpreting confidence level. The following exchange with Andy illustrates the type of confusion that Andy, Joe and Sandy appeared to experience as they reasoned through Student 2's interpretation.

> Interview 3: Gallup Poll:
>
> > **Andy:** Let's see so what's going on here with this one.…The first thing I'm thinking is that the 95% confident in the estimation process. It's not saying that. They could be 100% confident in the process. They could use that process every single time. They just they really like it. It says nothing about how confident they are in the process. The process works perfectly. It could be perfect for all we know. It's just that when you only sample a limited number of Oregonians you just don't get to know the truth. What you really wanted was to sample all of them, but that's just impractical so this is the whole point of the confidence interval. You're going to take a small, a smaller set that you could actually manage. You could be wrong. And so you've got to say how wrong are you likely to be given that, assuming that your process is perfect. Because the process has all kinds of problems, like the person that goes up and asks is really smelly, you know. And it's like, yeah sure tax me just get out of my face, you know. Or it's totally a spun question. It's like would you like to be bludgeoned if we don't tax you. I mean the way they spin the question, or the order in which it's asked. The process may have lots of faults, but it's not at issue.

250

I think that Andy's response to Student 2's interpretation suggests that he read "confidence in the estimation process" as questioning the way the survey was conducted, rather than the random error associated with collecting repeated samples. That is, Andy did not have an image of an idealized sampling experiment that could be repeated. Amanda suggested that the wording of Hypothetical Student 2's interpretation could be a source of confusion in the problem and it could account for why Andy, Joe and Sandy discredited Student 2's interpretation.

Summary of TA Reasoning about the Gallup Poll Task

Reflecting back on TAs' responses as a whole, there appears to be a continuum from which these TAs reasoned about the Gallup Poll Task, ranging from explicit and robust connections of confidence intervals to the distribution of sample statistics (Amanda), to less explicit and robust connections of confidence intervals to the distribution of sample statistics (Sandy), and finally to little or no connection to the distribution of sample statistics, but rather to confidence in the particular interval obtained in the sample (Andy, Joe, and Sam). It is not surprising that TAs would not explicitly discuss confidence level or an interpretation that entails repeated sampling in their interpretations, as often these ideas are implicitly understood (or assumed to be understood) in the conversation. Sandy's interpretation of confidence interval appeared to be at the other end of the spectrum from Amanda's interpretation in their initial responses. For example, based on TAs' initial utterances to the Gallup Poll

Task, Sandy, on the one hand, appeared to interpret the confidence interval as an interval that *would* contain μ. Amanda, on the other hand, appeared to interpret confidence intervals with an associated level of confidence in the sampling process. Yet, the problem with drawing this conclusion is that often statisticians have a tacit conceptualization of confidence level and repeated sampling, and the common vocabulary for discussing confidence intervals does not make explicit these tacit assumptions.

Using hypothetical student responses that did explicitly mention confidence level and repeated sampling forced these TAs to be more explicit about their meanings. In particular, Andy's, Joe's, and Sam's responses to the different hypothetical student predictions more strongly demonstrated that their conceptual images of confidence intervals were different than Amanda's and Sandy's. That is, Amanda and Sandy appeared to hold a frequency interpretation, whereas Andy, Joe, and Sam appeared to hold a subjective interpretation. Andy and Sam appeared generally surprised by the idea of repeated sampling in relation to the concept of confidence intervals. This is a rather surprising finding given the number of graduate statistics courses each of these TAs had and that they had each taught the first term of introductory statistics. In addition, Joe did not think that the concept of repeated sampling was an important idea to take away from the topic of confidence intervals, even if such an underlying image is statistically correct.

5.4 Conclusions

Taken together, the information about TAs' thinking gleaned from the interview conversations around the Unusual Sample and Gallup Poll Tasks suggest that there is a wide spectrum of ways in which these TAs conceptualize sampling and confidence interval problems. It is surprising that some of these TAs did not appear to see a connection between: (1) probability and sampling; (2) long-term relative frequencies of events and sample size; or, (3) repeated sampling and statistical inference. At least two TAs in this study (there may have been more in the survey, but there is not enough evidence to tell) reasoned using the Outcome Approach on the Unusual Sample Task. Three TAs in this study did not see a connection between repeated sampling and statistical inference. Frequency versus subjective approaches to probability lies at the heart of these differing interpretations. A frequency interpretation of probability relies on the long-term relative frequency of an event in order to assign probabilities to uncertain events. A subjective interpretation of probability relies on the degree of belief that a particular situation will occur once and only once to assign probabilities to uncertain events. These two different views of the underlying notion of probability result in different approaches to probability problems and how probability relates to ideas of sampling and statistical inference.

Further, one TA made pedagogical statements about how students should interpret confidence intervals that is in direct conflict with what the statistics education community actually advocates. That is, Joe suggested that confidence in the particular interval obtained in the sample was the important idea students needed to take away.

This is in direct conflict with statistics educators' call for developing an understanding of the connection between a distribution of sample statistics and statistical inference claims (Heid et al., 2005; Liu, 2004; Saldanha & Thompson, 2003). The pedagogical implications of TAs' content knowledge of sampling and statistical inference is discussed in Chapter 6.

Another compelling finding in this chapter is that each context posed a new situation to these TAs – in some cases a frequency view appeared dominant and in other cases a subjective view appeared dominant. For example, on the one hand, Amanda seemed to employ an Outcome Approach in reasoning about the Unusual Sample Task. That is, she appeared intent on answering the question of which sample *would* be unusual for that *particular* instance, rather than viewing that particular sample as one case of a larger set of possible samples. On the other hand, Amanda appeared more inclined toward a frequency approach as she reasoned through the Gallup Poll Task. She specifically discussed confidence level in relation to the concept of repeated sampling and what could be expected to happen in 95 out of 100 samples. Sandy also appeared more inclined toward an Outcome Approach as she reasoned through the Unusual Sample Task, and more inclined toward a repeated sampling perspective as she worked through the Gallup Poll Task. For Andy, the inclinations were reversed. In the Unusual Sample Task he appeared to take a frequency approach, raising the issue of the Law of Large Numbers and discussing the process by which larger samples tend to resemble the population parameter. Yet, in the Gallup Poll

Task, Andy did not find the student responses that entailed an image of repeated sampling to be coherent.

The fact that different types of reasoning were invoked by different contexts begs the question, why? Why would certain contexts lead a TA to reason from an image of repeated sampling, and other contexts lead a TA to reason about the outcome of the particular event? Of course, there are numerous variables at play here. For example, two possible variables that would influence TAs' reasoning are: (1) TAs' prior background and experiences working with these types of problems or more standard problems from introductory statistics curriculum (or higher-level statistics); and, (2) how TAs made sense of these prior experiences. I conjecture that TAs would be more likely to make a connection between repeated sampling and the distribution of sample statistics in relation to the Gallup Poll Task because this image is explicitly mentioned in introductory statistics texts, whereas using a distribution of sample statistics to reason about the Unusual Sample Task is less likely to be part of undergraduate or graduate coursework or textbook materials. In addition, I conjecture that most of these TAs may not be explicitly aware of how they are interpreting a probability statement or relating it to sampling and statistical inference problems, because often the focus in coursework is on calculations not on interpretations. Also, certain probability assumptions may be so implicit in instruction that students (including graduate students) are not picking up on them.

Understanding how TAs (and other graduate students) might reason in different contexts could be useful for professors as they think about how to support the kind of

reasoning desired of graduate students that has come to be accepted by the larger

statistics community. In addition, knowing how TAs reason in the different contexts

presented by these tasks sheds light on their statistical knowledge for teaching. Is that

knowledge developed enough so as to support the type and quality of undergraduate

learning for which statistics educators are calling? In the next chapter (Chapter 6) I

discuss the issue of TAs' statistical knowledge for teaching. In particular, I use the

information gleaned from this chapter to discuss how different ways in which TAs

reasoned through these tasks could either support or hinder undergraduate statistics

education.

CHAPTER 6

TAS' SUBJECT MATTER KNOWLEDGE AND KNOWLEDGE OF CONTENT

AND STUDENTS: IMPLICATIONS FOR TEACHING STATISTICS

The two previous chapters addressed the first goal of my study – an investigation of TAs' statistical content knowledge. The primary purpose of this chapter is to address the second goal of my study – an investigation of TAs' statistical knowledge for teaching. To be clear, TAs' statistical knowledge for teaching is deeply intertwined with TAs' statistical content knowledge. They are addressed in separate chapters for ease of presentation; however, I draw heavily on the findings of the two previous chapters in my discussion of TAs' statistical knowledge for teaching. In Chapter 4, I discussed the difficulties TAs appeared to experience in their attempts to connect a theoretical model to experimental data, and to make decisions using multiple sampling distributions created from an experiment. In Chapter 5, I discussed the different ways in which TAs interpreted sampling and statistical inference tasks. In this chapter, I argue that TAs' difficulties making decisions about or with experimental data, and their different interpretations of sampling and statistical inference tasks, have profound consequences on their statistical knowledge for teaching. In addition, I argue that in order to better support student learning, TAs should also have knowledge of students' statistical development and common conceptual hurdles. That is to say, in this chapter

I shift my analytical lens by focusing on some of the implications of the findings of Chapters 4 and 5 on TAs' statistical knowledge for teaching.

This chapter is presented in three main sections. In section 6.1, I review and make more explicit the framework for statistical knowledge for teaching that was developed in Chapter 2. This framework constitutes an end product of this study in that it emerged as a result of my data analysis; yet this framework is supported by existing mathematics and statistics education research literature and the melding of research on mathematical knowledge for teaching with the constructs of statistical literacy, thinking and reasoning. In section 6.2, I use my framework of statistical knowledge for teaching to highlight elements of TAs' statistical knowledge for teaching that need further development. In section 6.3, I address TAs' beliefs about teaching statistics and how students learn statistics. Although investigating TAs' beliefs was not part of my primary research goals, TAs provided information on their beliefs about teaching and learning during the interviews. Beliefs and knowledge are constructs that are intimately connected, making it impossible to ignore the issue of TAs' beliefs within this study. Thus, this chapter concludes by discussing the ways in which TAs' content knowledge and their own learning experiences likely influence their beliefs about how to teach statistics and how students learn statistics.

6.1 A Framework for TAs' Statistical Knowledge for Teaching

One of the primary goals of this study is to investigate TAs' statistical knowledge for teaching. Recall, from my review of the literature (Chapter 2), the work of Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski &

Mewborn, 2001; Ball & McDiarmid, 1990; Hill, Rowan & Ball, 2005) on mathematical knowledge for teaching, the work of Eisenhart et al. (1993) on teachers' procedural and conceptual knowledge, and the constructs of statistical literacy (Gal, 2004; Watson & Callingham, 2003) and statistical thinking (Pfannkuch & Wild, 2004). Melding these multiple constructs I lay a foundation for what statistical knowledge for teaching sampling and statistical inference topics may look like. After presenting this model in section 6.1.1, I discuss how it can be further specified in relation to the survey and interview tasks used in this study. I argue that TAs need substantial content knowledge, and knowledge of students' ways of thinking and common developmental paths in order to develop a robust understanding of statistics in their students.

6.1.1 Components of Statistical Knowledge for Teaching

The purpose of this section is to make a contribution to statistics education research by putting forth a framework that characterizes some critical components of statistical knowledge for teaching. It is certainly beyond the scope of this study to tackle a complete characterization of the necessary content knowledge for teaching introductory statistics. Yet, this study begins the development of a characterization of the statistical knowledge necessary for teaching by examining the key aspects of *content knowledge* and *knowledge of content and students* needed for teaching sampling and statistical inference topics.

259

Recall from Chapter 2 that Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Ball & McDiarmid, 1990; Hill, Rowan & Ball, 2005) frame the construct of mathematical knowledge for teaching into four components: (1) common content knowledge, (2) specialized content knowledge, (3) knowledge of content and students, and (4) knowledge of content and teaching[24]. Also recall that Eisenhart et al. use the constructs of procedural and conceptual knowledge to describe the types and qualities of teacher knowledge. These two frameworks serve as a useful starting point for an examination of statistical knowledge for teaching, yet they require refinement and revision in order to be useful for describing the type and quality of knowledge that statistics TAs (and teachers) should have. In the subsections that follow, I propose three components necessary for strong statistical knowledge for teaching – *statistical literacy*, *statistical thinking*, and *knowledge of content and students*. It is important to keep in mind that these three components are intimately intertwined, and it is not a useful exercise to neatly parse out TAs' knowledge into each distinct component. Rather these three constructs are useful for framing a discussion about what statistical knowledge for teaching looks like, developing methods for assessing TAs' statistical knowledge for teaching, and developing mentoring opportunities for TAs that enable them to improve their statistical knowledge for teaching.

Statistical Literacy as Common Content Knowledge

---

[24] Ball's fourth component, knowledge of content and teaching, is beyond the scope of this study.

Ball (2005) defines common content knowledge as "the mathematical knowledge and skill expected of any well-educated adult" (p. 13). Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Ball & McDiarmid, 1990; Hill, Rowan & Ball, 2005) research interests are in elementary school mathematics, so for them common content knowledge includes the ability to add, subtract, multiply, and divide real numbers. In the statistics education research community (Gal, 2003 & 2004; Ben-Zvi & Garfield, 2004; Watson & Callingham, 2003), the construct of *statistical literacy* serves as an illustration of common content knowledge. Recall from my review of the literature (Chapter 2) that statistical literacy is defined as the ability to be an educated consumer of statistics (Gal, 2004). That is, every educated adult in our society should be able to read, organize, interpret and critically evaluate statistical information presented by the media, internet sites, newspapers, and magazines (Gal, 2003 & 2004; Ben-Zvi & Garfield, 2004; Watson & Moritz, 2000; Watson & Callingham, 2003). In order to make sense of statistical information found in different media sources, the construct of statistical literacy includes conceptual and procedural knowledge of measures of center, conceptual knowledge of variability in statistical sampling, estimation skills, the ability to coordinate multiple attributes of a distribution, conceptual knowledge of a distribution of sample statistics, and the idea of repeated sampling. Watson and Moritz suggested a three-tiered hierarchy for statistical literacy: understanding basic statistical terminology, understanding terminology when it appears in social contexts, and the ability to question statistical claims in these contexts (p. 11). Watson and Moritz argue

261

that by the time a student graduates from high school he/she should be able to reason at the third level.

Statistical Thinking as Specialized Content Knowledge

Ball et al. (2005) define specialized content knowledge as "the mathematical knowledge and skill needed by teachers in their work and beyond that expected of any well-educated adult" (p. 22). Statistics TAs' (and teachers') knowledge should extend beyond statistical literacy skills if they are to teach statistics well. Statistics TAs need to be what Gal (2004) refers to as *producers* and *consumers* of statistics. That is, in addition to statistical literacy skills, statistics TAs need to know the formal procedures and tools of probability, sampling and statistical inference in order to engage in their own research and to teach. Additional content knowledge is also important as a means for understanding the "big ideas" and connections between and among statistical concepts, and as a means for articulating statistical explanations in the classroom. The "big ideas" can be described by Pfannkuch and Wild's (2004) construct of *statistical thinking*, which broadly describes the "statistical enquiry cycle, ranging from problem formulation to the communication of conclusions" (p.41). Shaughnessy (2007) suggests that statistical thinking can be thought of as normative thinking; that is the type of statistical thinking used by statisticians and accepted by the statistics community. Statistical thinking requires sufficiently deep knowledge of the procedures and concepts of a typical introductory statistics curriculum, including basic probability, sampling, and statistical inference and the relationships between these

concepts. In addition, statistical thinking requires knowledge of how to pose a well-specified research question or problem, design an experiment, collect data, analyze data and draw conclusions from the analysis. The omnipresence of variation must be kept in mind throughout this process, including possible sources of variability, and ways to control and/or quantify variability (Pfannkuch & Wild, 2004).

Knowledge of Content and Students

Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003) define the construct of *knowledge of content and students*. They suggest that this knowledge includes knowledge of common student misconceptions, identifying student errors and possible reasons for those errors, and knowing how students are likely to approach certain mathematical tasks. This component of statistical knowledge for teaching consists of TAs' (teachers') knowledge of students' statistical development within particular statistical domains. In my literature review (Chapter 2), I identified and compiled a list of common student misconceptions and stumbling blocks to statistical thinking[25] (see Figure 44).

---

[25] Recall that the difficulties highlighted in bold are particularly relevant to this study because of their prevalence in students, high school teachers, and the TAs in this study.

Figure 44: Common Difficulties in Understanding Sampling Concepts

Common Difficulties and Misconceptions in Students' Reasoning and Development of Sampling Concepts

- Difficulty with the concept of random sample
    - Difficulty distinguishing between the colloquial versus statistical use of the term random and the term sample
    - Difficulty recognizing sources of bias in sampling
- Difficulty with the added level of abstraction required for understanding sampling distributions
    - Difficulty with the difference between a distribution of a sample and the distribution of a collection of sample statistics
    - Difficulty with the distinction between empirical and theoretical sampling distributions (a sophisticated concept – difficulty documented in teachers (Heid et al., 2005))
- **Difficulty attending to multiple aspects of a distribution**
    - Overly focused on modes or other measures of center
    - Overly focused on variability or individual data points
    - Focus on shape – Difficulty making distinctions between the normal and other symmetric shaped distributions
- **Difficulty finding a balance between sample representativeness and sample variability**
    - **Difficulty understanding the role of sample size in sampling variability**
        - **Do not expect a difference in variability for different size samples or believe that large samples have more variability**
    - **Difficulty relating a long-term relative frequency view of probability to sampling and statistical inference problems**
- **Difficulty understanding the role of sampling in the creation confidence intervals**
    - **Difficulty conceptualizing confidence level and margin of error – maintaining an image of repeating the sampling process**
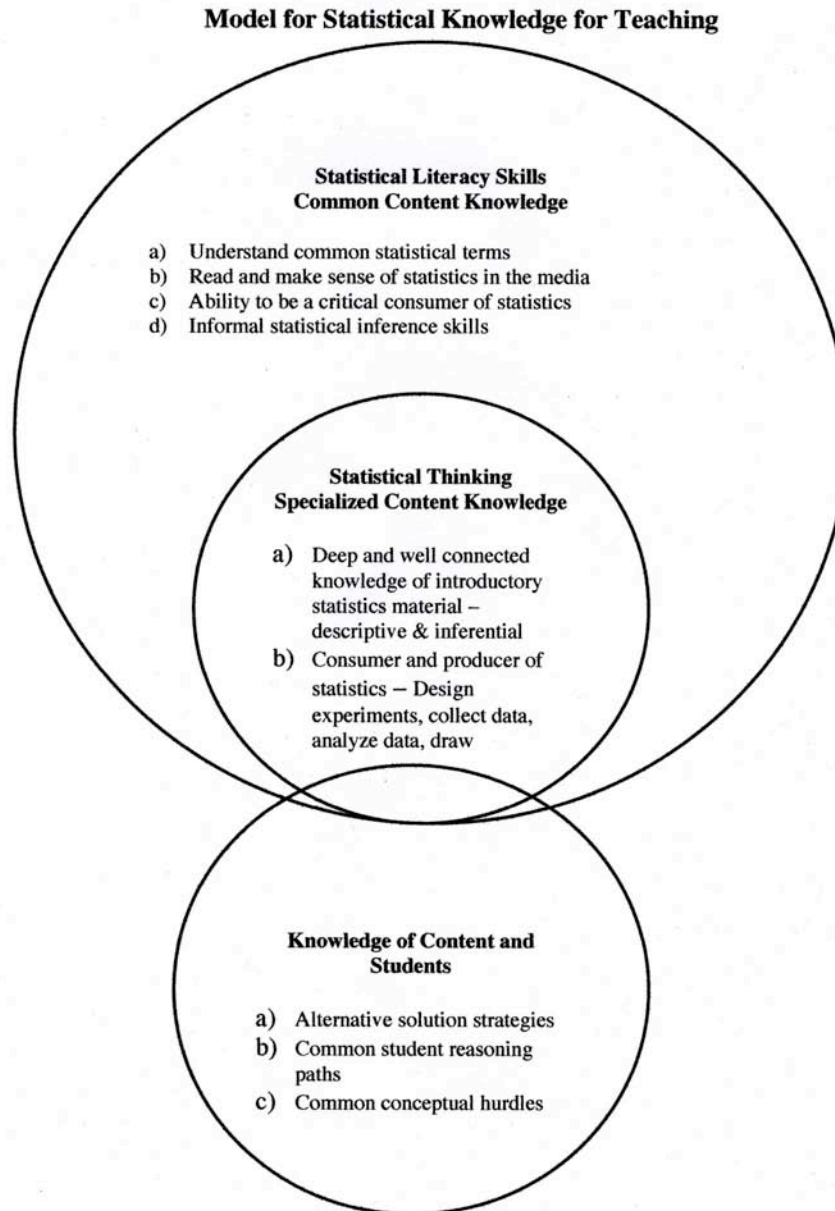- Difficulty with the concept of the Central Limit Theorem

I argue that TAs (and teachers) should have knowledge of the conceptual difficulties that students are likely to experience and the erroneous reasoning they may be likely to apply in order to provide instruction that can support student learning. I argue that if instruction employs student errors or common misconceptions as a place to begin statistical conversations in the classroom in such a way as to make explicit the subtle conceptual underpinnings of statistical topics, then students will be more likely to develop statistical literacy and thinking skills.

Interplay of the Components of Statistical Knowledge for Teaching

The Venn diagram, shown in Figure 45, provides a visual representation of my model for statistical knowledge for teaching. This model illustrates the key components of this knowledge, which I teased apart in the preceding paragraphs, yet expresses the overlap between these constructs. I mentioned places of overlap in Chapter 2 when I first discussed the constructs of *statistical literacy* and *statistical thinking*. For example, statistical thinking implies statistical literacy skills because it requires the ability to be both a *consumer* and a *producer* of statistics. Thus, statistical thinking is contained inside statistical literacy. In addition, certain aspects related to *knowledge of content and students* overlap with statistical literacy and statistical thinking. For instance, knowledge of alternative solution strategies is necessary for the working statistician who needs to see a problem from multiple vantage points or communicate findings to clients who do not have a statistics background.

Figure 45: Model of Statistical Knowledge for Teaching

Model for Statistical Knowledge for Teaching

Statistical Literacy Skills
Common Content Knowledge

a) Understand common statistical terms
b) Read and make sense of statistics in the media
c) Ability to be a critical consumer of statistics
d) Informal statistical inference skills

Statistical Thinking
Specialized Content Knowledge

a) Deep and well connected knowledge of introductory statistics material – descriptive & inferential
b) Consumer and producer of statistics – Design experiments, collect data, analyze data, draw

Knowledge of Content and Students

a) Alternative solution strategies
b) Common student reasoning paths
c) Common conceptual hurdles

I argue that it is also useful to think about these constructs as part of a connected

spiral. The evolution of TAs' knowledge in one component propels their knowledge in

the other components. For instance, as TAs gain a deeper understanding of the introductory statistics curriculum, they are likely to begin to see connections between different statistical ideas, which may, in turn, enhance their ability to make sense of statistical information presented in the media or in students' alternative solution strategies. This example could flow in the opposite direction as well; that is, sorting through students' questions or solution strategies may force a TA to think more deeply about a particular statistical topic, enabling the development of the TA's statistical content knowledge.

6.1.2 Applying my Framework of Statistical Knowledge for Teaching to the Interview Tasks

In this section I apply my framework to the interview tasks. In particular, I elaborate on the statistical knowledge necessary for teaching the statistical concepts present in the Prediction, Real/Fake, and Mystery Mixture Tasks. I follow with an elaboration of this framework for the Unusual Sample and Gallup Poll Tasks. Specifically, this section is presented in four parts. The first part examines statistical content knowledge – *statistical literacy* and *statistical thinking* skills – necessary for teaching sampling ideas related to the Prediction, Real/Fake, and Mystery Mixture Tasks. The second section examines *knowledge of content and students* in relation to sampling topics. The third section examines the statistical content knowledge necessary for teaching concepts of sample variability and the relationship between sampling and statistical inference as conceived of in the Unusual Sample and Gallup

Poll Tasks. The fourth section discusses knowledge of content and students in relation to sampling and statistical inference ideas.

Statistical Content Knowledge for Teaching: Prediction, Real/Fake, and Mystery Mixture Tasks

In Chapter 4, I provided a thorough conceptual analysis for the Prediction, Real/Fake, and Mystery Mixture Tasks. These analyses have the added advantage of serving as an aid to my framework for statistical knowledge for teaching in a sampling context. That is, the conceptual analyses help to specify the different components of statistical content knowledge necessary for teaching these tasks. Salient features that arose in my conceptual analysis for these three tasks are: (1) reasoning with measures of center, spread, or shape; (2) reasoning distributionally – informal or formal; and, (3) reasoning about experimental versus theoretical sampling distributions, particularly with reference to variability. I discussed each of these components in detail throughout Chapter 4. To remind the reader, I raise again a few of the key ideas.

- Reasoning with measures of center
  - Mode, median
  - Mean – Proportional reasoning, mean as balance point (center of mass), $\dfrac{\sum x}{n}$
  - Averaging the averages
- Reasoning with measures of spread
  - Range, Interquartile range

- Standard deviation, variance
  - Reasoning with measures of shape
    - Symmetric and skewed shapes – location of measures of center for different shaped distributions
  - Informal Distributional Reasoning
    - Combining one or more of center, shape, and spread to reason about a distribution
  - Formal Distributional Reasoning
    - Formal probability structures and assumptions for applying different probability models
  - Experimental versus Theoretical Sampling Distributions
    - Consideration of variability within and among sampling distributions
      - Sense of bounded variability or balance between variability and representativeness
      - Role of sample size in sampling variability
      - Image of repeatability of the sampling process

Indeed, knowledge of these features is important when reasoning throughout statistics, not just for these tasks. Knowledge of the informal features described above represent important components of statistical literacy because they entail the ability to (1) reason proportionally, (2) make sense of graphical information, (3) compare and contrast

graphs informally (by center, shape and spread arguments), and (4) understand the role that variability plays in experimental data. Each of these components is necessary for becoming an informed consumer of statistics. Knowledge of both the formal and informal features represent important components of statistical thinking, because taken together such knowledge structures provide a connection between the conceptual and procedural entailments, and the 'big ideas' of introductory statistics curriculum.

It is also important to note that the Prediction, Real/Fake and Mystery Mixture Tasks are somewhat open-ended tasks, to which there are no "right" or "wrong" answers; however, for each of these tasks there exist solutions that are certainly more likely than others. In the Real/Fake and Mystery Mixture Tasks, the experiment was performed via computer simulation (except in the case of the two fake graphs, which were manufactured by the researchers; Shaughnessy et al., 2004a&b). Thus, these tasks provided TAs an opportunity to investigate and make comparisons among experimental sampling distributions and to draw appropriate conclusions from those investigations. This focus on 'messy' experimental data and how to draw conclusions from such data are also key elements in the development of statistical literacy and statistical thinking because in real-world applications there can be numerous difficulties with the sampling process that influence how a statistician draws conclusions and/or how the lay public should interpret findings presented in the media.

TAs' Knowledge of Content and Students: Prediction, Real/Fake, and Mystery Mixture Tasks

Prior research (Shaughnessy et al., 2004a&b, 2005) suggests that middle and secondary students may reason about the Prediction, Real/Fake and Mystery Mixture Tasks by focusing on one attribute of the distribution (mode, shape, or spread), or reasoning additively – justifying their predictions on the fact that there are more red candies in the jar. In addition, some students become overly focused on measures of center and others become overly focused on variability. Other statistics educators have documented this tendency for students to struggle between the idea of sample representativeness and sample variability as well (Rubin et al., 1991). Figure 46, is a representation of the conceptual framework developed by Shaughnessy et al.[26].

---

[26] Recall that this research study was discussed in detail in Chapter 2

Figure 46: Conceptual Framework for Reasoning about Experimental Sampling

Distributions

Conceptual Framework of Shaughnessy et al. (2004, 2005)

Other (0)

Additive (1)

Transitional Stage
Mode (2)

Shape (2)   Variation (2)

Median/Mean (3)
Proportional
(Explicit connection between the sample and population %)

Distributional (4)
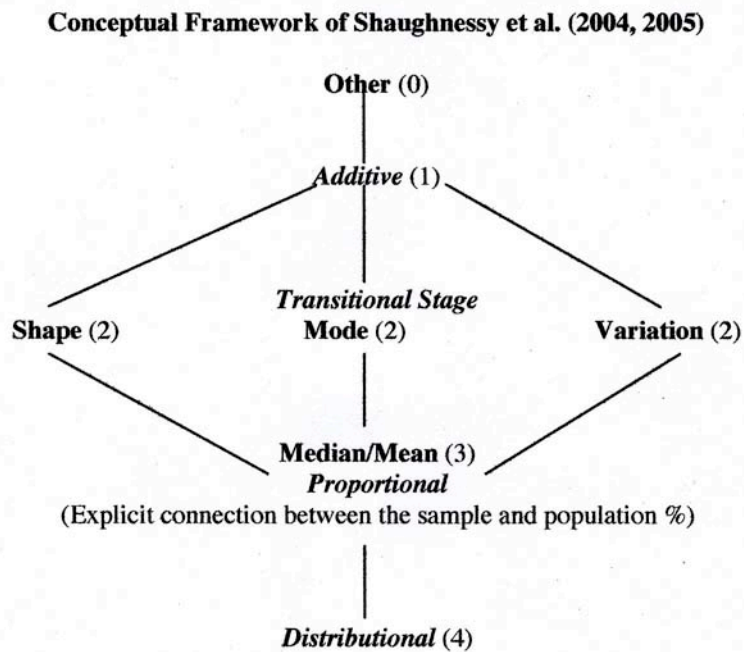
Past research investigating students' thinking and reasoning about these tasks

provide a foundation from which to examine TAs' knowledge of content and students.

The conceptual framework developed by Shaughnessy et al. (2004a&b, 2005)

suggests that TAs (and other teachers of statistics) need to be aware that students may

not have proportional reasoning skills or be able to apply them in this context.

Students may use additive reasoning, or focus on a single attribute of the distribution

such as shape, center (mode) or spread to make their predictions. TAs need to have

knowledge of student difficulties and the types of reasoning students are likely to

employ in order to provide instruction that can support student learning, using student

errors or common misconceptions as a place to begin statistical conversations in the classroom.

The research findings described above influenced the interview tasks for TAs. Recall that after TAs provided their own predictions in the Prediction Task, they were given the hypothetical predictions of two students and asked to comment on the reasonableness of those predictions (see Table 22).

Table 22

Number of Red Candies in Handfuls of 10 (Mixture: 750 red & 250 yellow)	Hypothetical Student 1's Predictions	Hypothetical Student 2's Predictions
0 red candies	0	1
1 red candies	0	1
2 red candies	0	1
3 red candies	0	1
4 red candies	0	2
5 red candies	5	3
6 red candies	9	4
7 red candies	15	17
8 red candies	18	18
9 red candies	3	1
10 red candies	0	1
TOTAL	50	50

Both hypothetical student predictions were designed to have reasonable and unreasonable qualities, relative to what normative models predict. Student 1's prediction was designed to be more reasonable than Student 2's prediction. However, Student 1's prediction was designed to be slightly too focused on the center, containing a substantial decline in outcomes from the left and right of the center at seven and eight red candies. Also, Student 1's prediction was designed with a more narrow range than might be expected. Student 2's prediction was designed to be much

too variable, containing an outcome in every possible location. In addition, Student 2's prediction was also too focused on the center at seven and eight red candies. The outcomes drop substantially on either side of the seven and eight red candy spots. I use the language "design" in reference to the hypothetical student responses, yet these responses came from actual student work and represented typical student responses (Shaughnessy, personal communication 2006).

In the Real/Fake Task, Shaughnessy et al. (2004a&b, 2005) noticed that many of the students relied on the shape, modes, or the most extreme values in order to justify whether or not a graph was 'real' or 'fake'. Just as I observed with TAs, many of the students in Shaughnessy's study either expected the shape of the experimental graphs to be ideal ('smooth') or they did not. In addition, the students either expected to have fewer or more outcomes with 9 and 10 red candies, or fewer or more outcomes with 2, 3, and 4 red candies. In the Mystery Mixture Task, Shaughnessy et al. observed that: (1) students experienced difficulty reasoning about multiple attributes of the distribution simultaneously; and, (2) many students, especially in the middle grades, tended to use modal averages.

Statistical Content Knowledge for Teaching: The Unusual Sample and Gallup Poll Tasks

In Chapter 5, I provided a thorough conceptual analysis for the Unusual Sample and Gallup Poll Tasks. The conceptual analyses helps to specify the different components of statistical content knowledge necessary for teaching these tasks. Salient

features that arose in my conceptual analysis for these two tasks are: (1) the image of repeating the sampling process; and, (2) interpreting sampling and statistical inference concepts with a long-term relative frequency perspective of probability. I discussed each of these components in detail throughout Chapter 5. To remind the reader, I raise again a few of the key ideas.

- Connecting the Law of Large Numbers to sampling contexts

  o The role of sample size in sampling variability

- The hypothetical repeating of the sampling process

  o Image of a distribution of sampling statistics

  o Knowledge that there is random, chance error in the sampling process and that this can be quantified with a long-term relative frequency view of probability

These key ideas are important for both an informal and formal understanding of the conceptual entailments of statistical inference. TAs should have knowledge of the conceptual underpinnings of each task, as well as how to apply formal theorems and techniques, such as the Law of Large Numbers and confidence interval formulas. It is also important to note that the Unusual Sample and Gallup Poll Tasks provided TAs an opportunity to make sense of statistical information in a contextual setting. In particular, the Gallup Poll Task represents the type of statistical information that adults in our society would encounter in different media sources. Thus, this task is also significant from a statistical literacy perspective.

TAs' Knowledge of Content and Students: The Unusual Sample and Gallup Poll Tasks

In Chapter 2, I discussed a number of research studies that indicate students' proclivities toward interpreting sampling and statistical inference problems, like the Unusual Sample and Gallup Poll Tasks, as pertaining to the particular sample, rather than interpreting such problems from the image of a distribution of sample statistics (Kahneman & Tversky, 1971,1972; Konold, 1989; Watson, 2004). Specifically, these past research studies indicate that students (K-12 and college) have a tendency to believe that large and small samples have the same amount of variability. That is, students fail to understand the role of sample size in sampling variability. In addition, students have trouble understanding the added level of abstraction in the concept of sampling distributions and fail to apply the image of a collection of sample statistics to sampling and statistical inference problems (Saldanha & Thompson, 2003).

Past research investigating K-12 and college students' thinking and reasoning about these tasks provide a foundation from which to examine TAs' knowledge of content and students. The research suggests that TAs (and other teachers of statistics) need to be aware that students: (1) may not have the necessary scaffolding in place to understand sampling distributions on a conceptual level; (2) may not apply an image of sampling distributions to sampling and statistical inference problems; and, (3) may not understand the relationship between sample size and sample variability. Such information about student reasoning could enable TAs to develop instruction that can support student learning of these particularly challenging concepts.

The statistics education research on content and students discussed here influenced the interview tasks and questions. After TAs provided their own response to the Unusual Sample Task, I asked them to describe any difficulties students might have with that task. For the Gallup Poll Task, I provided a number of hypothetical student responses and asked TAs to comment on the reasonableness of these responses. Some of the hypothetical responses were valid interpretations of confidence interval, but others where incoherent or suggested the confidence was in the particular interval from the sample and not in the sampling process (see Chapter 5 for the specific hypothetical student interpretations).

Summary

This section addressed key features of content knowledge – *statistical literacy* and *statistical thinking* – and *knowledge of content and students*. These features are important for TAs' statistical knowledge for teaching, and have been identified by other statistics education researchers as important for learners of statistics. For instance, Pfannkuch (2005) argues that the following components need to be developed in order for students to develop informal and formal understandings of inference:

1. Reasoning with measures of center

2. Distributional reasoning

3. Sampling reasoning

4. Drawing an acceptable conclusion based on informal inference (p. 275)

Pfannkuch also mentions that these concepts be developed using real data and a constant consideration of variation. Students must be able to reason with "messy" data sets and develop an understanding of bounded variability – that balance between sample representativeness and sample variability.

Pfannkuch (2005) suggests that by grade 10, students should have well-developed informal inference skills. That is, the ability to describe, interpret, and compare variability in data sets, rather than determine in a probabilistic sense how likely an event is to occur (p. 279). Pfannkuch's argument is important because it suggests that these informal inference concepts are necessary for developing students' statistical literacy skills and creating educated and informed consumers of statistics. With such instruction, students who desire to develop their statistical knowledge toward a more formal level of statistical inquiry are likely to have the underlying imagery and conceptual underpinnings in place to support the transition to formal statistical inference techniques. At this level, students learn the formal probability distributions and how to model different phenomenon using formal probability structures. Students should develop a deeper conceptualization that "random variation is described mathematically by probability" (Pfannkuch, 2005, p. 279), in the sense of repeated sampling. I believe that Pfannkuch's comment can be interpreted via understanding the balance between sample variability and representativeness, or the concept of

bounded variability discussed by Rubin et al. (1991) and Saldanha and Thompson (2003), respectively. That is, students need to have a sense for an expected variability among sample outcomes and among experimental sampling distributions. Students should learn more formalized concepts and procedures for discussing variability, such as standard deviation and variance. Students should learn procedures for finding confidence intervals and conducting hypothesis tests, and be able to interpret ideas of confidence level and margin of error.

Pfannkuch's (2005) pedagogical model for developing informal and formal knowledge of statistical inference in high school students has much in common with my model for necessary content knowledge for teaching introductory statistics. If these knowledge components set the stage for the development of statistical literacy skills and the pathway toward more formal understandings of statistical inference in students, it seems reasonable that these are exactly the type of knowledge structures that should already be well developed by TAs.

In sum, I argue that TAs should have knowledge of the salient features mentioned for the Prediction, Real/Fake, Mystery Mixture, Unusual Sample, and Gallup Poll Tasks, be able to apply that knowledge in novel settings, and draw appropriate conclusions. Such knowledge structures would certainly be an indication that TAs had well-developed statistical literacy and statistical thinking skills. In addition, TAs should be able to articulate this knowledge base verbally and to know something about students' development in this area. Taken together, such a skill set would entail necessary content knowledge for teaching these types of sampling concepts.

6.2 An Analysis of TAs' Statistical Knowledge for Teaching

Now that I have provided a framework for the statistical knowledge necessary for teaching, I use this framework to analyze TAs' statistical knowledge for teaching. In sections 6.2.1 and 6.2.2, I return to the themes presented in Chapters 4 and 5. These chapters discussed a number of key statistical ideas with which the TAs in this study had difficulty and they implied areas in which TAs had limited *statistical literacy* and *statistical thinking* skills. I discuss how TAs' difficulties with this content translate into limited statistical knowledge for teaching. In section 6.2.3, I discuss TAs' knowledge of *content and students*.

6.2.1 TAs' Statistical Content Knowledge: Prediction, Real/Fake and Mystery Mixture Tasks

Chapter 4 revealed that TAs' content knowledge, as evidenced by their responses to the Prediction, Real/Fake and Mystery Mixture Tasks, was not as robust as my framework for necessary statistical content knowledge for teaching recommends. Recall from Chapter 4 that TAs appeared to experience tension when making decisions based on data collected from an experiment. In particular, some key difficulties expressed by the TAs in this study include:

- Difficulty thinking about variability in the experimental data in a statistically coherent manner
  - Difficulty with notion of bounded variability – balance between variability and representativeness

- Overly focused on variability in the frequencies of the graphs

- Difficulty using a probability function to identify classes of graphs that are less likely or more likely to occur as a result of the experiment.

- Difficulty attending to multiple attributes of the distribution simultaneously or reasoning with multiple graphs

- Difficulty estimating the mean of a graphical distribution.

I argue that the key difficulties TAs experienced with these tasks is troubling from a pedagogical point of view. These TAs experienced some of the common difficulties researchers (Rubin et al., 1991; Shaughnessy et al. 2004a&b, 2005; Watson & Moritz, 2000) have identified in students (see Figure 44 in Section 6.1 for compilation of student difficulties). That TAs experienced these difficulties suggests that their statistical literacy and statistical thinking skills are not deeply developed. Of course, statistical thinking could be described as a lifetime goal; yet, some of the difficulties these TAs experienced are troubling given their level of experience in graduate statistics courses. Thus, TAs' difficulties with these tasks are problematic from a graduate statistics education point of view because their graduate coursework has not been successful at fine-tuning TAs' statistical content knowledge. It is also problematic from an undergraduate statistics education point of view because these TAs are teaching statistics to undergraduate students. I argue that if TAs' experience

the very problems that statistics educators are attempting to correct in students, then TAs are not well positioned to provide quality instruction in these areas. This conjecture is supported by research on teacher knowledge; for example, Eisenhart et al. (1993) and Even (1993) each made similar arguments based on their research of K-12 teachers.

To illustrate my position more clearly, I revisit the key difficulties TAs' experienced and discuss more specifically how these difficulties are problematic from a pedagogical perspective.

Real/Fake Task: Focus on Shape

Recall that both on the survey, and during the interview many TAs focused on the shapes of the graphs as a means for determining which graphs were 'real' and which were 'fake'. More in-depth questioning during the interviews revealed that TAs appeared focused on the variability in frequencies, in the sense of uneven frequencies as one reads the graph from left to right. In addition, the TAs I interviewed expressed their struggle with making conclusions about experimental data. For instance, recall Amanda's struggle as she attempted to make her real/fake justifications.

Interview 1: Real/Fake Task:

> **Amanda:** …I'm having a battle in my head about theoretically what I expect to happen, which would look like Graph 3, and reasonably in practice what I have seen happen.

TAs experienced tension that they could not always resolve between their expectations of the graphs of experimental data and their knowledge of the theoretical

model. The source of this tension appeared to be TAs' difficulty mitigating their expectations for variability versus representativeness in experimental data. With the exception of Joe, the TAs I interviewed did not expect an exact match between the graphs of experimental data and the theoretical model. Yet, they did not seem to have a strong statistical sense of how the graphs of experimental data should vary. These TAs focused on variability in frequencies, rather than statistical variability. This focus translates into TAs' attention on the single attribute of shape in the data. In particular, these TAs focused on the extent to which the shape of the experimental sampling distributions was smooth or uneven as a basis for making their real/fake identifications. I argue that this is problematic for teaching informal notions of variability. Given a data set, these TAs are likely able to compute the variance and standard deviation; yet, how can they provide a strong conceptual understanding to their students if they cannot discuss statistical variability in this context?

The work of Eisenhart et al. (1993) provides evidence of a teacher that becomes confused by the conceptual underpinnings of multiplication of decimals. As a result of her confusion, the teacher indicates that a visual approach to the problem would confuse students so she indicates that she would use a procedural approach when teaching decimal multiplication. Their finding suggests that when teachers do not understand the conceptual underpinnings of a topic they are more likely to teach procedures for finding correct solutions and avoid discussions about the concept or the reasons for a particular procedure. Thus, it seems reasonable to assume that if TAs do not have a strong conceptual understanding of variability or how to use their

knowledge of variability to reason about experimental data sets, then they are unlikely to implement instruction for developing this knowledge in students.

As I mentioned, Joe indicated that he believed Graph 3 was the most likely of the four graphs because he expected the graphs of experimental sampling distributions to match closely to the theoretical model.

Interview 1: Real/Fake Task:

> **Joe:** Yeah, yeah. If we assume the computer is a perfect random generator I would expect the most likely output to match this distribution that I have in my head.

Joe's conception that the experimental distribution should match that closely to the theoretical model for 50 trials of 10 is problematic because it suggests that perhaps Joe tends toward a belief that a collection of data from an experiment is *overly representative* of the population distribution. Recall that this perception has been documented in the literature (Rubin et al., 1991) in middle and secondary students. It seems likely that if Joe shares a common statistical misconception with middle and secondary students, that he will have difficulty developing instructional activities to combat this misconception.

Finally, the times that TAs did attend to other attributes of the distribution, they did not seem to coordinate the attributes, and their point of focus tended to be problematic. For example, Amanda did, at times, focus on variability in terms of the spread of the distribution; however, she appeared to expect more outcomes of two, three and four red candies than is likely to occur. Recall the following exchanges

where Amanda discussed her expectations for pulling out two, three and four red candies versus nine and ten red candies.

Interview 1: Prediction Task:

Interviewer: And what about 3 red candies in your handful or 10 red candies?
Amanda: I have a harder time imagining 10 red candies.
Interviewer: Okay. So you're thinking it's more likely I'm going to get 3 red candies?
Amanda: Umm, just in a visualizing sense. Yeah.

Interview 1: Real/Fake Task:

Interviewer: Okay, so why are Graphs 1 and 4 more likely to happen for you than the other two graphs?
Amanda: I think for the same reasons I felt like two and three were the fakes. Graph 2 doesn't have a lot of variation occurring…. We are incredibly heavy lumped in the 9 and 10 reds…. But in Graph 2, nothing is going on below 4, which makes me uncomfortable.
Interviewer: Why?
Amanda: Because I think it should, something… at least one observation below 4 should occur…. I've got nothing here [*in 2's, 3's and 4's*] and a lot going on at 9 and 10. And I would feel more comfortable. Watch this. This is just going to be awful. If I removed some off 9 and 10 and moved them over here [*to the 2, 3 and 4 red candy slots*] so that it looked more like Graph 3. The theoretical one [*laughs*] that I think is implausible…. I'm having a battle in my head about theoretically what I expect to happen, which would look like Graph 3, and reasonably in practice what I have seen happen.

These excerpts (as well as the other evidence provided in Chapter 4) suggest that Amanda's visualization of the distribution may not map well to the actual theoretical distribution, or that she does not have a strong conception of bounded variability, or both. Amanda appears to expect more outcomes of two, three, and four red candies than outcomes of nine or ten red candies. This discrepancy could certainly pose a problem in the classroom as Amanda attempts to draw upon her conceptual

285

knowledge/image of the situation to create learning experiences for her students. It would be difficult for Amanda to engage students in a didactically rich discussion around the real/fake graphs given that Amanda is torn between her image of the theoretical model, the actual theoretical model, and what she expects to see happen in an experimental situation. In addition, she does not appear comfortable using her statistical knowledge to resolve her tension. Again, there is support for my conjecture that limitations in TAs' statistical content knowledge translate to limitations in TAs' abilities to teach these concepts in the work of Eisenhart et al. (1993) and Even (1993).

Mystery Mixture Task: Coordinating Multiple Aspects of a Distribution and Reasoning with Multiple Experimental Distributions

Like the Real/Fake Task, to successfully solve the Mystery Mixture Task requires a deeper level of knowledge entailing both procedural and conceptual knowledge. The Mystery Mixture Task requires the ability to inspect experimental data and make inferences about the population parameter. No particular procedure or algorithm can be applied to determine the *exact* population parameter. Instead one must rely on their knowledge of distribution, and big picture ideas like: (1) measures of center, shape, and spread; and, (2) informal and/or formal probability distribution arguments. Recall that Amanda, Joe, and Sandy experienced difficulty using the four experimental sampling distributions in the Mystery Mixture Task to provide a point estimate (or interval estimate) for the proportion of red candies in the population. Amanda and Joe used modal averages to justify their predictions, and Sandy reasoned primarily by

shape, but only used one of the four graphs. These results are similar to the

observations by Shaughnessy et al. (2004b, 2005) in the middle and secondary

students they studied.

Amanda, Joe, and Sandy's reasoning on the Mystery Mixture Task suggests that

they had limited conceptual and procedural understanding in reasoning with multiple

experimental sampling distributions, which hindered their ability to solve this task

well. For example, Amanda and Joe did not use the median and/or mean to estimate

the population parameter; rather, they focused on the modes of each graph and found

the modal average. Whereas Joe never mentioned any other attributes of the

distributions in forming his estimate, the next exchange shows that Amanda at least

attempted to look at other aspects of the distribution.

Interview 1: Mystery Mixture Task:

> **Amanda:** I did momentarily make a conscious effort at examining the distributions across the other numbers. I can tell you right now that my brain froze up and I didn't know how to process that information so my brain went back to the concentration around the 2 and the 3. So I really did make an effort for a minute to consider 5, 6, and 7 [*red candies spots*] on these graphs, but it didn't feel. God I've never had to pick apart how I think about these things. But I instantly had a feeling of uncertainty about how to get any information about that, so I immediately went back to the mode.
> **Interviewer:** So kind of on a gut level you could go back to the modes and be able to say something.
> **Amanda:** Right, right.

Amanda abandoned her attempt to use multiple measures of center or other attributes

of the distribution. The previous excerpt suggests that she felt uncomfortable about

how to get useful statistical information from the distribution. While there is nothing

inherently wrong with taking the modal average for each of the four graphs to get an

estimate for the number of red candies in the jar, it is only one way of estimating that population parameter. Further, the median and/or mean provide better approximations to the population parameter.

Joe and Amanda certainly have enough knowledge to find their own solution to this problem and to provide a justification for their prediction. However, Joe and Amanda do not appear to have a deep conceptual knowledge necessary for solving this task. They were not able to use more sophisticated measures of center or coordinate aspects of the distribution to yield a better point estimate in this context. For example, neither seemed to recognize mean as a balance point and use that knowledge to estimate the means for each graph visually. In fact, recall that when Amanda was pushed to estimate the mean of each graph, she struggled. Amanda used the shape of the sampling distributions to discuss the location of the different measures of center, but did not seem to know how to use that information in helping her make her prediction. The difficulty Joe and Amanda experienced applying their knowledge of distributions, and, in particular, measures of center in a novel sampling situation are likely to hinder their abilities to use the task effectively in instruction. For example, with their limited knowledge, how would they engage their students in a discussion about applying past knowledge of measures of center to a new situation? There is support for this conjecture in the research of Eisenhart et al. (1993). Eisenhart et al. noted that the teacher, Ms. Daniels, they studied rarely taught for conceptual knowledge, despite her interest in doing so, and this was largely due to her own limited conceptual understanding of the material.

Recall that Sandy experienced difficulty reconciling how to use all four

experimental sampling distributions for estimating the population parameter.

Interview 1: Mystery Mixture Task:

> **Sandy:** So now I have no clue. So now it seems that you get it skewed to the right. So before I had 750 red. Okay, I had 750 red. I got it skewed to the left and 250 yellow. I'm just thinking that to see it like that I should have red smaller than yellow.
> **…**
> **Interviewer:** So would it be the same mix, but the reverse? 750 yellow and 250 red?
> **Sandy:** No, I'm not sure about that I cannot. I can see that for example here [*Sandy points to Graph 4*] it peaks at 2 and 3. If I reverse it, I knew that before it peaked at 7 and 8 [*in the real/fake task*]. Just by looking at this graph [*Sandy points to Graph 4 again*]. So it might be, it might be the same if I'm looking just here you know [*Sandy points to Graph 4*].

Sandy did not have sufficient conceptual knowledge of distribution to apply or transfer

that knowledge to a new situation. Other than her observation of the modes on Graph

4, she did not make an attempt to compute any measure of center for each of the

graphs or average a measure of center for the four graphs. She also appeared paralyzed

by having more sample information than she was used to and did not know how to

process all that information during the interview. Sandy expressed her dislike for this

problem because it did not have a set way for her to proceed.

Interview 1: Mystery Mixture Task:

> **Sandy:** …So how I approach problems is give me the information. I will apply the formulas. I will give you the answer. And I'm confident of doing that.

I believe that at the center of Sandy's dislike for the Mystery Mixture Task is that she

is used to attacking problems on a procedural level and she feels uncomfortable when

a problem does not fit within in a recognized format. Her limited conceptual knowledge of distribution, and her inability to apply her procedural knowledge for finding the median and mean of a data set in a novel setting is problematic for her own personal development and for developing deeper levels of knowledge in students. Specifically, Sandy would be taxed if she were to center a lesson around the Mystery Mixture Task because she would not have the ability to

- Solve this task confidently herself

- Recognize that more samples provide more information

- Use students' different ways of approaching the problem as a basis for class discussion

- Recognize acceptable alternative solution strategies

Summary

The TAs in this study experienced difficulty with these tasks because reasoning about experimental data using multiple attributes of the distribution and maintaining a statistical sense of bounded variability was problematic for them. This is obviously problematic for their statistical literacy and thinking, yet it is also troubling from a pedagogical perspective. Even's (1993) research suggests that teachers "pedagogical decisions – the questions they ask, activities they design, students' suggestions they follow – are based, in part, on their subject-matter knowledge" (p. 113). Thus, I also argue that the limitations in TAs' content knowledge as described in Chapters 4 and 5 (reviewed briefly again here) translate to insufficient knowledge for: (1) designing

quality statistical instruction; and, (2) making judgments about the reasonableness of student predictions/decisions. Specifically, I claim that the following set of questions would be appropriate questions to raise when thinking about how to solve the Prediction, Real/Fake and Mystery Mixture Tasks, and could serve as pedagogical tools for developing students' informal and formal statistical reasoning skills:

- How much variability can be expected from handful to handful and/or trial to trial?

- How many outcomes can be expected in the left and right ends of the distribution?

- What would an unusual number of low or high outcomes be?

- What interval contains 80-90% of the data?

- What shape can be expected for the experimental distributions?

- What would an unusual shape look like?

- What would constitute an unusual center?

- How do the mode, median and mean relate to each other in a skewed-left/skewed-right distribution?

- How can the means for each graph be visually estimated?

- How unusual is the dip at 3 red candies on Graph 3 in the Mystery Mixture Task?

These questions address the key features of statistical literacy and thinking mentioned at the outset of this chapter. Yet, these were not questions that the TAs in this study were sufficiently prepared to address for themselves or their students.

6.2.2 TAs' Statistical Content Knowledge: The Unusual Sample and Gallup Poll Tasks

Chapter 5 revealed that TAs' content knowledge, as evidenced by their responses to the Unusual Sample and Gallup Poll Tasks, was not as robust as my framework for necessary statistical content knowledge for teaching recommends. Recall in Chapter 5, I discussed the following key struggles among TA participants:

- Interpreting sampling and statistical inference problems from an image of repeated sampling rather than from an image of the particular sample collected
    - Relating probability from a long term relative frequency perspective to sampling and statistical inference problems
    - Image of a distribution of sample statistics
    - Random, chance error in the sampling process

The sources of TAs' difficulties with the Unusual Sample and Gallup Poll Tasks have been observed in secondary and tertiary students, as well as high school teachers (Kahneman & Tversky, 1971; Konold, 1989; Liu, 2004; Rubin et al., 1991; Saldanha & Thompson, 2003). Again, I contend that if TAs experience the same conceptual difficulties as beginning statistics students, then they are not well positioned to address

these developmental stumbling blocks in instruction. This contention is supported in the research literature, for example, Ball, Hill, and Bass (2005) and Eisenhart et al. (1993). I turn now to a more detailed discussion of the key difficulties TAs experienced during their work on the Unusual Sample and Gallup Poll Tasks and the ways in which this is problematic in instruction.

The Unusual Sample Task: Probability & the Image of Repeated Sampling

Recall that Amanda and Sandy both answered the Unusual Sample Task by suggesting that it was not possible to tell which sample *is likely* to be more unusual because the ratio of boy to girls in both school is the same. In Chapter 5, I argued that the way Amanda and Sandy approached the task suggested that it was probable they were answering the question which sample *will be* more unusual, and because they could not tell for sure, they marked both samples are equally likely to be the unusual sample. This, in and of itself, is problematic because it suggests, at least on an unconscious level, that Amanda and Sandy are interpreting probability in this context as pertaining to the particular sample obtained and not to an image of repeated sampling based on long-term relative frequencies. If Amanda and Sandy experience this difficulty then they will be unable to recognize this problematic thinking in their students. In addition, they will likely be ineffective at addressing the conceptual underpinning related to this task; that is, the role that sample size plays in sampling variability. If Amanda and Sandy do not see the importance of the distribution of sample statistics in relation to sampling tasks, then they are unlikely to highlight this

key feature of statistical inquiry during instruction. This is a fundamental concept, which is necessary for understanding variability in sampling distributions from different size samples, and in connecting sample size to statistical inference. Thus, it is important in the development of statistical literacy and thinking.

Sandy's content knowledge limitations are further reaching than Amanda's. Whereas Amanda recognized a relationship between the Unusual Sample Task and a Binomial probability distribution, Sandy did not see such a connection. Recall from Chapter 5 the following exchange, where Sandy argued that this task did not relate to probability.

Interview 1: Unusual Sample Task:

Interviewer: Some of the TAs taking this survey answered exactly how you did, but some answered that the small sample from the country school…
Sandy: No I see no reason for that.
Interviewer: Well, one TA used the Law of Large Numbers for his reason.
Sandy: The Law of Large Numbers.... I don't really see that. I don't see the relationship because this problem is about collecting a sample, you know….Yes. Yes, you have 50/50 boys/girls, but it's not a problem about umm. …It's not a similar problem. I don't think. Not from my point of view. You just go and you sample some people randomly, randomly.… Because like here [*referring to the coin problem*] you are doing the same thing over and over again. While here [*referring to the sampling problem of boys and girls*] you go and take a sample, a random sample. You're not supposed to go again and go again.

Sandy's knowledge appeared compartmentalized and she did not see a connection between sampling and probability. Even (1993) noticed this same "compartmentalization phenomenon" in her work studying teachers' conceptions of functions. Even suggested that inconsistencies in teachers' conceptions of function

might be explained by conflicting schemes in their cognitive structures, which are stimulated by different contexts. In Sandy's case, she did not demonstrate flexibility to conceptualize her sample of 20 children as 20 trials and her sample of 50 children as 50 trials, where each trial could result in a boy or girl. Again, this is problematic from a pedagogical point of view because Sandy would not recognize the application of the Law of Large Numbers as a valid approach to the problem. In addition, she would lack the ability to develop those connections in her students.

The Gallup Poll Task: Probability & the Image of Repeated Sampling

The Gallup Poll Task requires an understanding of the procedures for finding a confidence interval, margin of error and confidence level, and the conceptual underpinnings of how to interpret the information the interval provides and what the confidence level means. The procedural aspects of this task did not appear problematic for these TAs. Amanda, Sandy, Joe, and Andy were all comfortable calculating the confidence interval using the point estimate and margin of error. Joe and Andy were unfamiliar with the procedure for calculating the confidence level for the poll, yet they seemed confident that if they were given the formula they would be able to "figure it out". It was the conceptual underpinnings for how to interpret confidence level that appeared problematic for Andy and Joe, not the procedures for calculating confidence intervals. The work of Eisenhart et al. (1993) suggests that if TAs do not have a strong conceptual understanding of confidence level, then it is likely to translate into a TAs' inability to develop students' conceptual understanding of confidence intervals.

Recall that in Chapter 5 I discussed how Andy's initial responses to the Gallup Poll Task suggested that he interpreted the confidence level as being related to the particular interval calculated from the sample statistic. Also recall that Andy appeared rather surprised by the hypothetical student responses that suggested an image of repeated sampling. The following exchange provides strong evidence that Andy did not view confidence level as being related to an image of the sampling distribution and long-term relative frequency interpretation of probability.

Interview 3: Gallup Poll Task:

Interviewer: Okay, another student said we don't know if the interval 32% to 40% contains the true percentage of voters that believe an income tax is necessary, but if we sample 100 times about 94% of those times the interval would capture the true percentage of voters.
Andy: Wait a second. [***Long pause***]. Yes, it's true that we don't know that this interval contains the true percentage. If we are going to sample 100 times, 94% of those will be in this interval. [***Andy reads the second part of the student's response aloud***]….Yeah that's not what this confidence interval says. It doesn't say a thing about re-sampling. It doesn't imply re-sampling. It talks about the margin of error and I'm going to stick with my definition [***laughs***], that it's not related to the re-sampling or the hypothetical re-sampling of it.
…
Interviewer: Could there also be an implied idea of repeated sampling?
Andy: There could be, but I've never understood it to be that way. Maybe it is and I've always just misinterpreted it.

Like Andy, Joe also did not suggest an interpretation of confidence level consistent with an image of repeating the sampling process and the collection of a distribution of sample statistics.

In addition, as Joe negotiated through the hypothetical student responses, he presented an argument for why a view of confidence level consistent with a subjective

interpretation made more pedagogical sense. That is, Joe argued that it was important for students to interpret the confidence level as being related to how likely the *particular* interval obtained from the sample was to capture the population parameter.

Interview 3: Gallup Poll Task:

> **Joe:** …This [*Student A*] I think is, in a certain sense, a better answer because it captures the meaning of what's going on here.
> **Interviewer:** Student A does?
> **Joe:** Yeah. And like the realistic meaning when I read a newspaper and I'm thinking about what statistics reported mean….The useful way to think about it is how many people think a sales tax is necessary. What does this statistic mean? 36% think it is, with a margin of error of ±4%.
> **…**This [*Student A's response*] captures the meaning and the interpretation that's important contextually for an educated layperson.

Joe and Andy *did not* appear to interpret confidence intervals in a manner consistent with a long-term relative frequency perspective. That is, they did not appear to have an image of the distribution of sample statistics in relation to the level of confidence. Their image of the confidence as pertaining to the particular sample is problematic from a pedagogical point of view because they are likely to: (1) provide their students with a subjective interpretation of confidence; (2) neglect the importance of a distribution of sample statistics and the image of repeated sampling in their presentations of the material; (3) neglect to make connections between probability with a long-term relative frequency perspective and statistical inference; and, (4) dismiss alternative student interpretations based on repeated sampling.

Summary

On the one hand, TAs should know that the interval they calculate does not necessarily capture the true population parameter. That is, there is some level of error associated with the confidence interval, and the level of error is with respect to the random error associated with the sampling process, not with respect to the particular interval computed. TAs should have knowledge that the confidence level relates to confidence in the sampling process, meaning if the experiment were repeated over and over again (a large number of times) the confidence level is the percentage of intervals that capture the population parameter. For statistics educators (Chance, delMas, & Garfield, 2004; Konold, 1989; Liu, 2004; Pfannkuch, 2005; Rubin et al., 1991; Saldanha & Thompson, 2003) this conceptual knowledge of confidence intervals is fundamental for statistical literacy skills and serves as a basis for understanding more formal statistical inference procedures. The importance of this repeated sampling perspective in the statistics education community translates to an important didactic theme to develop in instruction. It seems logical that these TAs were not in a position to support this key knowledge component in students given that it did not appear to be an explicit part of their mental schemas. Other researchers have argued that limited subject matter knowledge translates to limited mathematical knowledge for teaching (Ball, Hill, & Bass, 2005; Eisenhart et al, 1993; Even, 1993). In addition, TAs' reasoning appeared to be situational and depended on their own understanding of a

298

particular context. TAs situation dependent reasoning has been observed by Even (1993) in her work with secondary teachers. Thus, TAs did not appear to be cognizant of how they were interpreting these ideas, or even that this debate in how to interpret probabilities existed in the statistics community.

On the other hand, it is not surprising that the image of a distribution of sample statistics was not explicitly (or even implicitly) part of TAs' mental schema because these ideas (1) are not trivial, (2) may not have been explicit in their own undergraduate and/or graduate statistics education, and (3) tend to be underemphasized in introductory statistics texts. There is certainly evidence for my second point from Sandy. Sandy mentioned numerous times that at the time of our interviews she was enrolled in a graduate statistics course (for non-statistics majors[27]) that essentially presented the introductory statistics curriculum in a condensed manner and from a more sophisticated mathematical perspective[28]. The following exchange shows one instance of how this class impacted Sandy's conceptual knowledge of confidence intervals.

Interview 3: Gallup Poll:

> **Sandy:** …Because the more textbook interpretation is like in 95, if you repeat the sampling process 95% of the times, you'll capture in your confidence interval the true parameter, population parameter. We just had this discussion in the stats class I'm taking, and I learned of the repeating the sampling process, but our professor said that it is okay to use this, 'I am 95% confident that' [*points to the student's interpretation of confidence level that we just read*].

---

[27] The course was a graduate statistics course for engineers and computer science majors.
[28] Essentially, the course was calculus based.

> **Interviewer:** That the interval you just got captures the population proportion?
> **Sandy:** Yeah, yeah. And I believe that maybe this is not very exact.

In this exchange Sandy articulates the difference between having confidence in the particular interval versus confidence in the sampling process. Sandy was in her last term of a graduate program in statistics and had over 11 graduate statistics courses, yet it was not until this statistics course for non-majors that she encountered the idea of repeated sampling in a way that began to connect for her.

In addition, it is worth noting that in the previous exchange Sandy also indicated that the professor communicated either interpretation was "okay". The fact that professors often use, as shorthand, an expression that conveys confidence in the particular interval is problematic both for undergraduate and graduate education because while a statistics professor may have an implicit image of repeated sampling in his/her mental scheme, it is likely to go unnoticed by students. The same pedagogical problem resides in undergraduate statistics texts, which often ignore or downplay the idea of repeated sampling in their discussions of confidence intervals. Take for instance Larson and Faber's (2000) *Elementary Statistics: Picturing the World and More*, which fails to mention the image of repeated sampling in the interpretation of confidence level. I suggest that the implicit image of repeated sampling in relation to confidence intervals makes it difficult for both graduate and undergraduate students to develop the explicit imagery of repeated sampling that statistics educators advocate so strongly for.

6.2.3 Knowledge of Content and Students

Using the construct of *knowledge of content and students,* developed by Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Ball & McDiarmid, 1990; Hill, Rowan & Ball, 2005), as the third component in my framework, and the compilation of common difficulties and misconceptions students (and teachers) experience, which I synthesized in Chapter 2 (see Figure 44, Section 6.1), I examine TAs' knowledge of content and students. I begin with the Prediction, Real/Fake, and Mystery Mixture Tasks and then follow with the Unusual Sample and Gallup Poll Tasks. I conclude with some methodological comments about particular tasks that seemed better suited to eliciting information from TAs about their knowledge of content and students.

Analysis of TAs' Knowledge of Content and Students: Prediction, Real/Fake, and Mystery Mixture Tasks

Table 23 provides an overview of the common student difficulties to which the TAs in this study attended. Notice that TAs only demonstrated knowledge of students' tendencies to be overly focused on either measures of center or measures of variability, although this is not to say that they did not express other types of pedagogical knowledge.

Table 23: Knowledge of Content and Students – Prediction, Real/Fake, &

Mystery Mixture Tasks

Common Developmental Hurdles for Students in Prediction, Real/Fake, and Mystery Mixture Tasks	Evidence of TAs' Knowledge of Content and Students			
	Amanda	**Sandy**	**Joe**	**Andy**
Difficulty attending to multiple attributes of the distribution (1) Overly focused on measures of center – especially modes (2) Overly focused on measures of variability (3) Focus primarily on shape	✓ (1)	✓ (2)	✓ (2)	✓ (2)
Difficulty with the balance between representativeness and variability & recognizing the role of sample size in variability of samples and sampling distributions				
Difficulty recognizing the distinction between a sample of observations and a sample of statistics				

TAs' *knowledge of content and students*, in regards to students' tendencies to be

overly focused on either measures of center or variability, was expressed during their

responses to the hypothetical student predictions from the Prediction Task. I did not

find evidence of TAs' knowledge of content and students during our conversations

about the Real/Fake and Mystery Mixture Tasks. I suggest two reasons for why TAs

did not demonstrate knowledge of content and students for these two tasks. First, it is

possible that the task design did not elicit responses from TAs that could provide

evidence of their knowledge. In the Prediction Task, TAs were provided with

hypothetical student responses and asked to discuss those specific responses, whereas

in the Real/Fake and Mystery Mixture Tasks TAs were asked to discuss the types of

difficulties students would be likely to experience with the task. I discuss this methodological consideration in more detail in the section summary. The second possible reason that TAs did not express knowledge of content and students for these two tasks is that TAs experienced some of the same common difficulties as students studying statistics. I turn now to an analysis of the knowledge of *content and students* for which TAs expressed some evidence.

Amanda was the only TA out of five interviewed who explicitly recognized that Hypothetical Student 1's predictions seemed overly representative at the expense of sample variability (see Appendix for Hypothetical Student 1's prediction).

Interview 1: Prediction Task

> **Interviewer:** For Student 1, you marked their prediction as reasonable. Can you explain why?
> **Amanda:** Okay, so this indicates 0 out of the 50 will result in 0 through 4 red candies and then 0 for 10's. It's too extreme for me. It's all so heavily lumped in the 7 and 8's, and right around the 7 and 8's. So it's too concentrated around 7 and 8. We have nothing going on for 0 through 4 [*red candies*] and 10 [*red candies*] either that's why I don't think it's a great prediction. The reason I think it's reasonable is because we are talking about an introductory stats student and clearly this student is able to identify we're going to have heavy draws, that a lot of our 50 draws are going to result in 7 or 8 red candies. Then you even see a little downward dispersion from 7 to 8, not a great dispersion, but. I would be perfectly happy if an introductory stats student was able to put that much of the puzzle together.

The previous excerpt reveals that Amanda's criteria for a reasonable prediction from introductory statistics students were that they recognize the appropriate location of the center, place most of the outcomes near the center, and have some amount of dispersion around the center. Yet, Amanda maintained that Student 1's prediction

needed improvement in the distribution around the center because his/her prediction

was too narrowly focused at the center – "heavily lumped in the 7 and 8's and right

around the 7 and 8's".

Amanda also noticed that Hypothetical Student 2's prediction was overly

concentrated around the center, but again she said that she would consider it a

reasonable prediction.

Interview 1: Prediction Task

> **Amanda:** Well again, we have what I consider a very nice concentration
> around 7 and 8, which makes me happy that they were able to get that
> much out of the problem. Then it bothers me a little bit that it drops
> dramatically. We have 18 counts for 8 red candies, and 1 count for 9 red
> candies. And I'm not very comfortable with that…. It just so happens that I
> don't feel comfortable with the drop they've created from the 8 red candies
> to the 9 red candies…. I feel like the concentration at 7 and 8 red candies
> being 17 and 18 is too high.

On the one hand, Amanda noticed students' tendencies to focus too much on measures

of center. I observed Amanda's consideration of this tendency again during the third

interview when she discussed her knowledge of content and students with reference to

statistical inference concepts (this discussion is postponed until the end of this

section). On the other hand, Amanda did not appear to be bothered by the extreme

variability in Hypothetical Student 2's prediction. I asked her directly about the wide

range of Student 2's prediction when it did not come up naturally in her utterances.

Interview 1: Prediction Task

> **Interviewer:** Okay, is there anything, because you talked before about how
> unlikely it is for you to get 0 or 1 red, and I know they only have 1 here,
> but is there anything troubling to you about them having that 1 there?

Amanda: No. No, that doesn't trouble me at all. Because I debated 0 or 1 red candies, so it doesn't bother me at all that they chose to put 1 in there…. And if you can develop the notion, even just vaguely, that I'm going to have most of my, it's going to be much more likely that its going to 7 and 8 then it is to get 0 or 10. So I want most of my draws to go into 7 and 8. And I want the least of them to go into 0 and 10…. Then 0 and 10, say these are going to be the least common or the most extreme, and then filling in the gaps between what you expect to happen a lot and what you expect to happen a little.

Again, there is evidence in this excerpt that Amanda did not see a problem with outcomes in the 0 or 1 red candies spot because she struggled with these placements as well. Amanda's utterances about visualizing the situation with the majority of outcomes at 7 and 8 red candies and the fewest outcomes at 0 and 10 red candies provides further evidence supporting my earlier conjecture that Amanda's concept image is of a distribution with a wider range and more outcomes below five red candies than the actual distribution for this context. Given that Amanda struggled with her own visualization of the Prediction Task context versus the outcomes predicted by the binomial probability distribution, it is not surprising that Amanda did not recognize this as problematic in student work.

The three previous excerpts with Amanda not only provide evidence of her knowledge of content and students with respect to some students' tendencies to become overly focused on centers, but it also provides evidence of her thinking about introductory statistics students' developmental maturity. For example, at the end of the first excerpt, Amanda suggests that Student 1's prediction is reasonable and that she "would be perfectly happy if an introductory stats student was able to put that much of the puzzle together". In the second excerpt, Amanda again suggests that Student 2's

prediction is reasonable and says that it "makes me happy that they were able to get that much out of the problem". These statements provide some insight into Amanda's ideas about what introductory statistics students are capable of doing. These ideas influence how she assesses student work. There is little evidence to suggest that Amanda's notions about students' capabilities are grounded in any educational theory, at least she did not demonstrate that knowledge during the interview conversations. Rather, Amanda's judgments about students' abilities are likely based on her own beliefs, perhaps grounded in her own learning experiences, and/or her observations of students during her classroom teaching experiences[29].

Joe, Sandy, and Andy also discussed the variability of Student 2's prediction. Joe and Sandy pointed out the unreasonableness of placing a "1" in each of the values 0 through 3 red candies. Joe appeared primarily focused on the extreme left value of 0 red candies. The next exchange shows that Joe recognized how unlikely this student's prediction was and why a student might place at least one value in each outcome.

Interview 1: Prediction Task

> **Joe:** The chance of getting 0 [*red*] is, in a handful of 10 if I do 50 grabs, of getting that once, is phenomenally low.
> **Interviewer:** Okay, so this student has an unrealistic idea?
> **Joe:** Yeah, and because there is something in each block, I look at that and I mean. What I think is that the student said there has to be one everywhere, and I know we're going to get a peak right here, so I think that's good. But just because we have a tail doesn't mean we are expecting to see anything in it. I saw some bad mathematics there.

---

[29] Beliefs are discussed in more detail in Section 6.3 of this chapter.

The last two highlighted utterances provide some evidence that Joe had knowledge

that students might be tempted to place an outcome in every spot just because the

possibility for such an outcome existed, regardless of how unlikely it was. This

suggests that perhaps Joe could, at least in this situation, view the problem from the

perspective of a novice student.

Sandy recognized that Hypothetical Student 2's predictions containing at least one

outcome in every place were unlikely, especially in the left end of the distribution.

Interview 1: Prediction Task

> **Sandy:** This one is totally, totally off. I look here, so the probability of getting in a draw 0, 1, 2 or 3 [*red candies*] is very small. So I believe it is very hard, even if you have the jar of candies near you and you want to conduct this experiment. I believe it's very hard to get this situation that this student came up with.
> **Interviewer:** So what do you think the student is thinking? Why would they do this?
> **Sandy:** My only answer is that in fact they could have no clue what is going on. You know.

Sandy last utterance in this excerpt could suggest that she does not place much value

on knowledge students may bring with them to class – thinking that students are

"clueless", or it could suggest that she did not have any sense for why a student might

have this difficulty. The latter interpretation is not surprising in the sense that Sandy

did not do this experiment with her class and she is not studying developments in

statistics education for her degree program. However, it is unfortunate from a

statistical knowledge for teaching perspective because this particular stumbling block

has been well documented in the research literature (Reading & Shaughnessy, 2004;

Rubin et al., 1991), and combating this difficulty is relevant to statistical literacy

efforts. Specifically, aiming to increase people's awareness of how they think about the probabilities of highly unlikely events; for example, their chances for winning the lottery.

Andy's response to Hypothetical Student 2's prediction is distinctly different than the other TAs I interviewed in the sense that the context of the classroom setting appeared to be essential in order for him to place judgment on students' work. In the next exchange, Andy initially responded that Student 2's prediction was reasonable because the student had a general sense that the data would be clumped between seven and eight red candies and then would disperse to the left and right. Yet, as our conversation continued it became clear that the purpose of the instructional setting determined how Andy decided upon his criteria for judging the reasonableness of a student's prediction.

Interview 1: Prediction Task

> **Andy:** …Here their tails are thin. The tail [***pointing to 0 and 1 red***], if you look at this as representing the tail it's thin. It's not 0, but then for a lot of students the difference between 0 and 1 is small. So maybe their sort of mentally rounding up just a little bit, and they didn't round up to two or three or anything like that.
> **Interviewer:** Okay, so if this was a two here [***for the number of handfuls with 1 red candy***]. So I'd like to push you to see if this is reasonable? What would be unreasonable for you?
> **Andy:** In that case, it would depend on what I expect them to understand. If this is a year later, then I'm not expecting them to remember the exact distribution or run a calculation. I'm just expecting them to remember the feel of it. And this is real fat, most likely here. Here at 7.5 they've really straddled it. They've really got a sense that 7.5 is a big deal. They know something is happening right there, that's where the action is. Then the action dies off out here. It's just sort of noise, they got 0's or 1.
> **Interviewer:** But you're not bothered by the fact that this is now spread all the way out to the tail [in Student 2]?

Andy: I'm not worried about that, because if you mentally picture the distribution it's got a big fat center and thin tails. That thinness to a student at this level, they're not thinking about 10^{-20}. They're not going to go to that level.

Interviewer: Do we want them to be?

Andy: That depends on where they need to be focused. Do I want them to be able to look at this problem and have a gut feel for it or are we looking for more precision out of them?…Yeah, if the focus is precision and likelihood, then a 1 is ridiculous. A mental sketch, yeah I think that's reasonable…. Again with the caveat that we haven't had the discussion of small numbers. Because that's its own discussion. To talk about how unlikely something is. If it's my goal to say hey don't get screwed by the lottery then we are going to have that discussion. Then if this becomes this business of putting a one out here at this extreme then that becomes totally unacceptable.

Notice that Andy's criteria for a reasonable student prediction are based on his instructional goals. For instance, Andy initially speaks about the situation in terms of his expectations of a student a year after completing his statistics course. In this case, he would expect them to remember the general feel for the problem and thus, providing a prediction that roughly agrees in center, shape, and spread is sufficient. However, Andy indicates that if the goal of instruction was on precision or how to be an educated gambler, then he agrees that Student 2's predictions of 0, 1, and 2 red candies each happening one time is unacceptable.

That Andy's criteria for judging student work depended on his goals for instruction and were highly situational is evidence of general pedagogical knowledge. In part the previous excerpt shows that Andy distinguished between short and long term goals of instruction, something that the other TAs in this study did not appear to do during the interviews. It is unclear what might account for this difference in

response. Andy had approximately the same number of years of teaching experience as the other TAs, yet he seemed to show more thought in his instructional goals.

In the Real/Fake and Mystery Mixture Tasks, I could not find evidence of TAs' knowledge of content and students that mapped to findings in the statistics education research community. However, TAs did express beliefs about learning and teaching in responding to the pedagogical questions related to these tasks. For example, Sandy did not believe that she would use the Real/Fake Task in the classroom. The following exchange reveals why.

Interview 1: Real/Fake Task

> **Sandy:** I believe it's confusing. And then what's the message? You want your students to get some concepts clear, or just to know that everything is possible and everything can be interpreted in different ways? It's not white. It's not black. It's like today is pink [*laughs*].

Sandy's utterances indicate that she did not think this task would be a good teaching tool because it did not have a "correct" answer and was too open to interpretation. Joe expressed a similar sentiment as Sandy, in that he also thought the task was too confusing to give to students. Sandy's and Joe's suggestions that the task is too confusing is an indication of their own confusion in how to approach the task. Eisenhart et al. (1993) noticed the same preference in their case study teacher, Ms. Daniels, for avoiding problems that she found difficult herself. In addition, the previous excerpt with Sandy suggests that her view of teaching and learning is through a clear step-by-step approach, rather than through messy applied problems.

However, Sandy did indicate a preference for letting students actually perform the experiment with the added caveat that she had enough class time.

Other times TAs expressed an indication that there could be many places a student would become stuck on a particular task, but could not specify a particular source of confusion. For instance, on the Mystery Mixture Task, Andy indicated that students would express their lack of knowledge for how to begin such a problem.

Interview 2: Mystery Mixture Task

> **Andy:** Oh God they'd have all kinds of difficulties. Let's see how many kinds of difficulties might people have? Let me count the ways. They are going to have…. I'm going to get questions like, or statements like: 'I just don't know where to begin'. 'I just don't get it'. So ask probing questions and the answer is still, 'I just don't get it'. Then you have to plan some more things. But I think for any person actually physically doing the experiment and then even though you get these protests of 'I just don't get it', physically having done it they at least have it in their bones, right…. Whereas if they haven't done the experiment then they just have no clues, they are nowhere near it.

Andy did not discuss specific problems that might occur or types of reasoning students might employ to solve the task. Instead, he suggested general difficulties, like students who "just don't know where to begin". Yet, there is evidence of strong pedagogical knowledge in his utterances. For instance, Andy advocates for having students do the experiment because he suggests that such work provides students a basis for thinking about the task that they would not otherwise have. Andy suggests that after having done the experiment then when a student does not know where to begin, there is a foundation in place for him to start asking probing questions. These utterances suggest that Andy was thinking about instruction with students as the focal point, rather than

some outside curriculum as dictated by a textbook or departmental syllabus. Andy was also the only TA who did not add the caveat, "if there is time", after suggesting he would do the experiment with his class, perhaps an indication that hands-on experiments were more important instructional tools for him than the other TAs.

Analysis of TAs' Knowledge of Content and Students: The Unusual Sample and Gallup Poll Tasks

Table 24 shows where TAs provided explicit evidence of knowledge of content and students. Notice that, in general, I did not find evidence of TAs' knowledge of content and students in this context. I found no evidence in my analysis of TAs' discussion of content and students during the Unusual Sample Task that would indicate they had knowledge of the common developmental hurdles shown in Table 24. However, I did find some evidence that Amanda and Sandy were cognizant of student difficulties interpreting confidence intervals during our discussion of the Gallup Poll Task. I also found evidence of other types of pedagogical knowledge.

Table 24: Knowledge of Content and Students – Unusual Sample & Gallup Poll Tasks

Common Developmental Hurdles for Students in the Unusual Sample & Gallup Poll Tasks	Evidence of TAs' Knowledge of Content and Students			
	Amanda	Sandy	Joe	Andy
Difficulty understanding the role of sample size in sampling variability (1) Do not expect a difference in variability for different size samples or believe large samples are more variable (2) Difficulty relating a long-term relative frequency view of probability to sampling and statistical inference tasks				
Difficulty understanding the role of sampling in the creation of confidence intervals (1) Do not maintain an image of repeated sampling	✓ (1)	✓ (1)		
Difficulty recognizing the distinction between a sample of observations and a sample of statistics				

On the one hand, the lack of evidence of TAs' knowledge of content and students on the Unusual Sample Task as compared to the Gallup Poll Task could be partly due to the difference in task methodology. In the Unusual Sample Task, I asked TAs to speculate on the types of difficulties students would be likely to encounter, whereas in the Gallup Poll Task I provided specific hypothetical student interpretations for TAs to consider. On the other hand, the lack of evidence could also be due to the fact that some of these TAs experienced the same difficulties that the statistics education research community has documented in students. This factor has been proposed by Even (1993) in her work with teachers. For example, it is not surprising that Amanda

and Sandy did not display knowledge of student difficulties with the relationship between sample size and sample variability in relation to the Unusual Sample Task because they displayed this same difficulty. Likewise it is not surprising that Joe and Andy displayed no evidence of student difficulties with interpreting confidence intervals with an image of repeated sampling because, like many students, Joe and Andy appeared to express confidence in the particular interval obtained in the sample, rather than confidence in the sampling process. That is, Andy and Joe *did not* maintain an image of repeated sampling in their interpretation of confidence level. Further, Andy and Joe appeared dismayed by hypothetical student responses that suggested repeated sampling.

Amanda and Sandy did maintain an image of repeating the sampling process in their interpretations of confidence level and they appeared to recognize that this interpretation might be difficult for students to grasp. For example, in the next exchange I asked Amanda about the reasonableness of Student A's interpretation and she began a conversation of how subtle the idea of confidence level is. That Amanda recognized that this is a common interpretation of confidence interval for beginning statistics students is significant evidence of her knowledge of content and students in this context.

Interview 3: Gallup Poll Task

> **Interviewer:** So the first student says that the margin of error being 4% means that between 32% and 40% of all Oregon voters believe an income tax is necessary.

Amanda: I think that. I think that this is a very common interpretation. I don't think it's fully accurate, but I think conceptually it's how we process information, like 32% plus or minus 4%.

Interviewer: What do you think is inaccurate about it?

Amanda: That with a certain level of confidence the true proportion will be between 32 and 40%, but the truth of the matter is that the true proportion is either in this interval or its not. And this way of talking about it I think is very common. I think it's even okay because it's the way people process the information and I think it makes the most sense. … What does a confidence interval really mean? And it takes a while. And I think I'm pretty lenient with my students about it to, about the way they phrase it. About, I try to get them in the habit of when they say it, you know let's assume 95% confidence, with 95% confidence we can assume the true proportion would be between 32 and 40%, but that we're actually talking about 95% of all samples would capture the true proportion and either our sample did or it did not.

Notice that Amanda decided this student's interpretation is acceptable, but incomplete. Amanda appeared to recognize that Student A's interpretation is a natural way in which people tend to process the idea of confidence intervals. Here Amanda distinguished between what she considered an acceptable common interpretation for a novice or layperson and a more formal understanding of confidence interval. In the previous section Amanda provided an indication of what she believes that students are capable of in relation to the Prediction Task. In the previous excerpt there is again evidence of Amanda expressing her beliefs about her students' abilities. Amanda suggested that beginning statistics students are not likely to have the mathematical maturity to understand the subtleties in how confidence intervals are interpreted.

Sandy suggested that interpreting the confidence interval from the perspective of the particular sample is acceptable if the student knows the meaning behind the

interpretation is in relation to repeating the sampling process. The following exchange

shows Sandy's response to Hypothetical Student 1's interpretation of confidence level.

Interview 3: Gallup Poll Task

> **Interviewer:** So now I have two different students' interpretations of confidence level and so I would like to know if you agree or disagree with these students' interpretations. So the first one is: 95% confidence level means that you can be 95% confident that the particular interval you found in the survey captures the population proportion.
> **Sandy:** …Because the more textbook interpretation is like in 95, if you repeat the sampling process, 95% of the times you'll capture in your confidence interval the true parameter, population parameter. We just had this discussion in the stats class I'm taking, and I learned of the repeating the sampling process, but our professor said that it is okay to use this, 'I am 95% confident that' [*points to the student's interpretation of confidence level that we just read*]. And I believe that maybe this is not very exact.
> **Interviewer:** To say this is not very exact?
> **Sandy:** I believe so, but…
> **Interviewer:** So why, what is lacking in this?
> **Sandy:** I am 95% confident, because, ahh, you miss saying that you have to sample many, many times. You know…. I'm thinking if someone knows what it means and uses this I'm 95% confident [*Student 1's interpretation*] it's okay because he knows what is behind. But I'm afraid a student might take this literally without knowing that in fact there is something else, like you have to repeat the process many, many times and 95% of the times you capture this parameter.

Sandy indicated that Student 1's interpretation is acceptable, but imprecise. I have

previously noted that at the time of the interview Sandy was enrolled in a graduate

statistics course for non-statistics majors. Sandy cited that her professor used the same

interpretation as shorthand as evidence that Student 1's interpretation is acceptable.

Sandy's suggestion that the common shorthand notation is acceptable because her

professor used that interpretation is consistent with Lortie's (1975) idea of

"apprenticeship of observation". Having observed her own professor interpreting a

similar situation to the Gallup Poll Task in the same manner as Student 1 provided Sandy justification for using that response with her own students. That is, Sandy cited her professor's use of a similar interpretation to that of Student 1's as a basis for justifying the use of that interpretation in the classroom. In the last highlighted utterance Sandy indicated that missing from this interpretation is the idea of repeating the sampling process. Sandy suggested that as long as students understand that concept there is no reason to state it each time. Yet, Sandy also recognized the potential for an introductory statistics student to miss this interpretation.

When TAs did not supply knowledge of common student difficulties or developmental paths as indicated by the research literature, they tended to display some general pedagogical knowledge and/or beliefs about the task or how students learn. For instance, Joe suggested that the Unusual Sample Task was too wordy and that this would confuse students.

Interview 1: The Unusual Sample Task

Joe: I mean, yeah, all the language. And I'm sure if I saw this tape again I would be confusing, conflating the words that I want to use for the concepts.

Amanda expressed a similar concern about students' abilities to weed out the extraneous information.

Interview 1: The Unusual Sample Task

Amanda: In my experience there are too many words.
Interviewer: Too many words for students?
Amanda: Yes, they get bogged down in, there's so much information here that is not really necessary to the question at hand. Or maybe it is. I don't know. Maybe city schools versus rural schools makes a difference. I'm not

provided with knowledge that it's making any difference here. That would be extraneous. There are so many pieces of information here that you don't need in order to evaluate this question. Not that I would be willing to extract any of it or remove any of it before showing it because in a real research situation there's going to be lots of information that you don't need and you have to be able to pick out what you do need but I can tell you that my students, from my experience, would get hung up on all this other information. It would totally freak them out, and I feel like specifically, and this is awful, it's not formulaic enough a question.

Although Amanda and Joe did not display specific knowledge of common student difficulties, they did show general pedagogical knowledge in this instance in the sense that they recognized that students might have trouble solving any type of task that was overly wordy, or, as Amanda noted, "not formulaic enough". This knowledge could also be considered some general mathematical knowledge for teaching in that they recognized that students tend to have trouble with applications or novel mathematical tasks.

Summary

In general, I did not find evidence that TAs had substantial knowledge of common student difficulties or developmental stages. I have already alluded to two potential reasons for the paucity of TA knowledge in this area. I raise those reasons again here and add a third reason. First, these TAs experienced some of the same difficulties that have been observed in middle, secondary, and tertiary students. I documented specific difficulties TAs experienced with the interview tasks in Chapters 4 and 5, which included: (1) difficulty reasoning with experimental data sets, (2) recognizing appropriate variability in experimental sampling distributions, (3) recognizing the role

of sample size in sample variability, and (4) applying the image of a distribution of sample statistics to sampling and statistical inference problems. It seems logical that if these TAs experience difficulty with these concepts, then they will not recognize them as problematic in students. This argument has been made by other researchers as well, for instance, Ball, Hill and Bass (2005) with elementary mathematics teachers and Even (1993) with secondary mathematics teachers. Certainly this is problematic from a pedagogical standpoint because these TAs lacked a deeper content knowledge from which to think about statistics and how to teach it.

Second, these TAs were still relatively new teachers. The TAs in my study had all been teaching for at least three years, but some of them had only taught the first term of introductory statistics one time – Joe and Andy, for instance. Amanda and Sandy had taught the first and second terms of introductory statistics more than twice. It is plausible to suspect that it takes some amount of time teaching a course before noticing prevalent problematic reasoning or developmental stages in students.

Third, there is a methodological consideration. The Prediction and Gallup Poll Tasks allowed me to achieve some insight into TAs' knowledge of content and students, whereas the Real/Fake, Mystery Mixture, and Unusual Sample Tasks did not. There was a fundamental difference in the task design that could explain the difference in the outcomes. For the Prediction and Gallup Poll Tasks, there were hypothetical student responses that followed after TAs gave their own responses. TAs were asked to discuss the specific hypothetical student responses, which seemed to set up a situation in which TAs addressed a particular strength or inadequacy in the student's

319

work. In contrast, after TAs provided their own solutions to the Real/Fake, Mystery Mixture, and Unusual Sample Tasks, I asked TAs how they thought students might respond to these tasks and what kinds of difficulties students might experience. The open wording of these questions might explain why TAs tended to discuss general problems that they believed students might have, their personal opinions of the task, and/or their beliefs about teaching and student learning, rather than specifying particular difficulties. The fact that TAs did express their beliefs about teaching and student learning is both interesting and compelling, which is why the last section of this chapter addresses the issue of TAs' beliefs.

Before addressing the issue of TAs' beliefs there is one final point that deserves mention here. Although TAs did not display specific knowledge of content and students in relation to sampling concepts, they did display other forms of pedagogical knowledge. For example, Andy displayed pedagogical knowledge when he discussed his goals for instruction as determining how he would judge the reasonable of a student's prediction to the Prediction Task. That is, Andy suggested that his purpose for teaching a particular topic and his desired outcomes in terms of student knowledge determined the focus of classroom instruction and set the criteria for how his students are assessed. Joe and Amanda provide a second illustration of TAs' pedagogical knowledge. Both Joe and Amanda suggested students would struggle with word problem tasks, which are set in context and contain extraneous information. Joe and Amanda noted that students experience difficulties with word problems because they do not know how to determine which information is critical and which information is

extraneous. That TAs displayed other forms of pedagogical knowledge is important because it shows that TAs think about their students' learning. Using TAs' pedagogical knowledge could be a good place to begin professional development and the work toward improving TAs' knowledge of content and students.

6.3 TAs' Beliefs about Teaching and Student Learning

The primary goal of this study was to investigate TAs' statistical knowledge for teaching, not to investigate TAs' beliefs about the teaching and learning of statistics. However, the constructs of knowledge and beliefs are intimately intertwined. During the interviews, TAs' beliefs about teaching and student learning surfaced. The purpose of this final section is to address the issue of TAs' beliefs within this study.

To begin, I argue that TAs' prior experiences learning mathematics influences the way in which they conduct their own classes. Teachers' prior learning experiences is not addressed by Eisenhart et al. (1993) as a limiting factor in teachers' ability and/or interest in teaching for conceptual knowledge versus procedural knowledge; yet, this factor appeared important in shaping TAs' methods of instruction and their beliefs about learning. Ball and McDiarmid (1990) and Even (1993) discussed the tendency of teachers to teach in a similar manner to how they were taught. I speculate that these TAs primarily had experience in mathematics courses where the focus was on procedural knowledge first and only secondarily on conceptual knowledge. Further, I argue that these prior educational experiences are likely to result in TAs espousing similar pedagogical practices. This final section of Chapter 6 is organized into two sections. In the first section, I discuss in more detail TAs' prior mathematical

experiences and their characterizations of effective teaching. In the second section, I discuss the possible influence of TAs' prior experiences on their beliefs about teaching and learning.

6.3.1 TAs' Prior Mathematical Experiences & Characterizations of Effective Teaching

Each of the TAs in my study suggested that defining characteristics of effective teaching included the ability to clearly explain difficult mathematical processes, clearly connect concepts to processes, and to challenge students. From this defining characteristic, it seemed clear that the TAs in this study concerned themselves with imparting both procedural and conceptual knowledge to their students, yet their responses to many of the interview tasks suggested they were more focused on procedural knowledge. In this section, I use conversations from Amanda and Sandy to highlight my discussion on TAs' beliefs about teaching and learning. I choose Amanda and Sandy because they provided the clearest articulation of their beliefs about teaching and learning. The interview data and information gleaned from Andy, Joe, and Sam did not provide discrepant or corroborating evidence. They are not included in the discussion simply because they did not articulate their beliefs to the same extent as Sandy and Amanda, making it difficult to draw any conclusions.

In the following exchange, Amanda describes two influential mathematics teachers she had as an undergraduate. It is worth noting in this exchange that Amanda

emphasized qualities such as her teachers being challenging but fair, and her teachers'

clear explanations as positive qualities in these teachers.

> **Interviewer:** Could you describe an influential teacher that you had as an undergraduate and how you think that teacher might characterize good teaching?
> **Amanda:** Yes, 'Dr. S', or equally so 'Dr. O'. I have to split them because I try to, in my own teaching, I don't try to embody them, but I notice things that I do that they did. 'Dr. S' was a terrifying teacher that everybody just warned you about how scary he was. I remember sitting down for my first exam with that man and felt like I had been sitting in the wrong class all term and not recognizing any of the material on the test. He was very demanding and very challenging. He certainly didn't give you anything that looked like it came out of the book. But he was also incredibly fair. I remember getting a 42 on that test and it was a B, so he was incredibly fair as well. His focus was driven by not did you get a problem correct, but how did you approach the problem. You know what tools did you pull out to approach the problem. Did you conceptualize it correctly, not did you arrive at a correct answer. …. And Dr. O, Dr. O was, I swear the man opened his mouth and God spoke through him because I understood every word the man ever said about anything. I had almost zero need for my textbooks. I never really had questions. Everything he ever said was so clear and so easy to understand and I pray every day that I'm explaining things as well as he could explain them. …. I think they would think good teaching is being very challenging and demanding excellence from your students, or demanding thoughtfulness from your students.

Amanda suggested that Dr. S was concerned with the process of problem solving

more than obtaining a correct answer, but this does not necessarily indicate a focus on

the conceptual underpinnings of a problem. Dr. S might have focused on the

conceptual underpinnings, or the steps in the process, or both. Amanda alludes to

conceptual understanding as his focus, but there is not enough information from this

excerpt to make this conclusion. Amanda admired the clarity of Dr. O's explanations,

which is of course an important quality for effective teaching. However, Amanda's

utterances about never having to use her textbook, and never having questions, raises

some questions for me about her image of effective teaching. In particular, Amanda's comments lead me to wonder if she sees the teacher as the center of the classroom, with the job of imparting wisdom on her students through crystal clear lectures, rather than teacher as facilitator, whose job it is to pose questions for students to answer and with which to struggle? There is not enough information in Amanda's utterances about good teaching to infer much about her philosophy of teaching.

Sandy indicated that she did not have a high school[30], undergraduate, or graduate (for her first master's degree in mathematics) teacher that she liked, or that she felt embodied good teaching. Sandy suggested that her prior teachers of mathematics were only interested in putting theorems and their proofs on the board and students were left to solve problems on their own. In the following exchange, Sandy explained what mathematics was like for her in high school.

> **Sandy:** The manuals back home, so you have each section is mostly theory, no examples. And then you have at the end of each section tons of problems. And they will just come, paint on the board the theory and then you are left alone to figure out how to solve the problems. So it's very hard, very hard. So we used to have some books only with problems and sometimes problem solutions for those problems, and that's how I got it. Basically I was learning by myself math in high school. I was reading books, you know and reading solution books, reading solutions to problems to learn how to solve problems.

Sandy suggested that her undergraduate and first graduate school experiences were much the same. Sandy's utterances in this excerpt suggest to me that she learned mathematics by teaching herself the procedures using textbooks that contained example problems with solutions. Sandy expressed her frustration that the teachers she

---

[30] Sandy went to a private high school in her home country that specialized in mathematics and science.

had did not connect the theory or concepts with the problems she was required to

solve. Sandy did mention an influential teacher, 'Dr. J', from her graduate studies

program in statistics. The following exchange indicates that Sandy liked that he

challenged students, but that he explained the concepts in a way that made sense to

her.

> **Sandy:** He tries to explain the concepts. He doesn't just paint the concept. Okay, so this is it. He tries to explain and interpret the concepts and that's what makes him good. He challenges, I believe he challenges the students. I remember when I first started and I took the intro to math stat. I was challenged to do the homework for that class, and the exams they were challenging. I believe it's a very good balance. I believe he gives, but he also asks for back, and if you miss, it's points off.

Notice that Amanda's and Sandy's descriptions of the characteristics of effective

teaching were similar; that is, good teachers present material clearly, connect concepts

and processes, and challenge students. Yet, each of these features could mean different

things to different people. Thus, it is difficult to infer too much from such a brief

description of these TAs' mathematical learning experiences. However, Amanda's and

Sandy's descriptions of good teaching, coupled with some of their interview responses

about student learning, may provide some initial clues into their beliefs about teaching

and learning.

6.3.2 The Influence of TAs' Prior Experiences on their Beliefs about Teaching and Learning

I conjecture that Amanda's and Sandy's prior learning experiences, which seemed

focused primarily on procedures and processes, lead to a theory of learning whereby

procedural skills and processes are considered the first steps to learning mathematics, and that over time, and after much practice, the conceptual knowledge filters in. The fact that Amanda and Sandy espoused beliefs about teaching and learning based on their prior learning experiences is not unique. For example, Lortie (1975), and Nathan and Koedinger (2000) observed that teachers' beliefs about teaching and learning are often rooted in their prior learning experiences.

During our discussion about good teaching, Amanda did not articulate how she learned mathematics, but in her responses to some of my interview tasks on student thinking, Amanda provided both an explicit and implicit view of her beliefs of how students learn statistics. In particular, she discussed how she sees conceptual knowledge following from procedural knowledge because this is the way she learned. For example, during our discussion of the Unusual Sample Task, Amanda suggested that students would have difficulty with the task because it was not procedural enough.

Interview 1: Unusual Sample Task:

> **Amanda:** …I feel like specifically, and this is awful, it's not formulaic enough a question.

As the conversation continued, Amanda expressed the importance of both procedural and conceptual knowledge and a view for how she believed students' mathematical development might occur.

Interview 1: Unusual Sample Task:

> **Interviewer:** With like a definite way to proceed, that they can go?
> **Amanda:** Yes. Its not even, I don't even. I'm questioning myself every step of the way. Did I think about that correctly? Did I assume too much information? Did…there are a million places that I feel like I'm getting off

track. And those aren't even questions that they know to ask themselves. So for, especially for a 243, 244 student, it's such a delicate level because its stuff that is so, it's so foreign. It's not like any math they have ever seen before. I find they need something more formulaic to work through. Especially with statistics it takes a lot of time before the full concepts come together, and I think it's important to get them thinking. To introduce ideas that are just generating conversation about what's going on here, but I also think it's important to go through a rigorous process, and that, that starts to lead to an understanding of why something is occurring.

In the highlighted utterances Amanda suggested that students need something more formulaic to work through when new mathematical ideas are presented to them, but that conversations centered about statistical concepts were important toward building conceptual knowledge. I realized that Amanda was emphasizing both mastery of skills and an understanding of underlying relationships, but that she appeared to be focused on mastery of skills first. I asked her about how she envisioned the conceptual knowledge building from the procedural knowledge.

Interview 1: Unusual Sample Task:

Interviewer: The process of repeating the steps over and over again is what leads to…
Amanda: Yes, but that could just be because that's what happened for me in all of my mathematical experience. I just repeat something over and over and over. Even if I don't know why I'm doing something, and then as I grow through my mathematical understanding I begin to understand. Oh this is why I'm doing step 8. This is why I'm doing step 9, and why they fit together. And this is why I get a good result, or a result I need, or whatever. And I think for a lot of statistical ideas it's too much to ask a student things that are really abstract. Or maybe they are not really abstract, but just that foreign. I don't know if it's fair to ask them to explain something on a conceptual level. They can explain something on a function level: this is my variable of interest, I'm looking for the probability of something occurring, I'm going to use this formula because I know its binomial because there was only two outcomes. And you repeat this process over and over again, and the conceptual part of the why its occurring the way that it is starts to build.

Amanda was quite explicit in her belief that mathematical understanding on a

conceptual level grows from mathematical understanding on a procedural level.

Amanda recognized that this is how her own mathematical development occurred and

so it seemed reasonable to her that this might also be how her students' mathematical

development might occur. Thus, Amanda recognized a need for discussing the

conceptual entailments of a particular statistical idea in the classroom, but she focused

first on processes and did not necessarily believe that students should be held

accountable for conceptual understanding in the same way that they should be for

procedural understanding. As Amanda discussed her experiences teaching the Central

Limit Theorem, her view of learning mathematics (based on her own mathematical

experiences) continued to present itself.

Interview 2: Terminology - Central Limit Theorem

> **Amanda:** I don't think, any. I don't think I've ever had many students really get a grip on the Central Limit Theorem. I think it's very, it's a very lofty idea to them. It's probably the only place in 243/244 where you dig out some theory. Everything else is like here are the rules, follow the rules. But we rely on the Central Limit Theorem and I don't know how other teachers address it, but I, I always remind them we can do this because of the Central Limit Theorem, and I'll do a quick refresher on the Central Limit Theorem. But I think it's too, I think it's too abstract for them to really get a grip on. Like in terms of mathematical maturity they're probably at the level where they just want the rules. They want to follow the rules. I don't think any of my students ever really process what the Central Limit Theorem means.

Here again, Amanda made a distinction between teaching statistical rules versus the

concepts and the underlying structure of those rules. Amanda recognized that the

Central Limit Theorem is more abstract than most of the other material presented in

328

introductory statistics texts, and she decidedly believed that most students at that level are only able to make sense of the procedural aspects of the material.

Through these exchanges it seems reasonable to infer that Amanda has developed a theory for how students learn grounded in her own learning experiences. That theory appears to inform her that students learn concepts through the repetition of procedures and so teaching procedures is what leads to students' eventual conceptual understanding. My conversations with Amanda indicated that she focuses first on rules and procedures with her students, discusses the "how" and "whys", but does not hold students accountable for deeper conceptual knowledge. Even (1993) also noticed this tendency for teachers to teach students rules by which they can get the right answers without needing to understand the concepts.

Sandy certainly suggested that she learned mathematics by working on problems over and over again and comparing her work to solutions manuals when available. There is also evidence from the interviews that Sandy stressed procedural knowledge over conceptual knowledge in her teaching practices. For example, during the first interview I asked Sandy if she would use a question like the Unusual Sample Task with her own class, and as the following excerpt reveals, she decidedly would not.

Interview 1: Unusual Sample Task:

> **Sandy:** No, I don't believe so. Because I believe when someone learns it's good to have problems with straight answers and not dubious questions like this one, which makes people think okay what is the right answer. I believe when I want to learn something I want to learn based on clear questions. You know, such that I get the concepts and later on if I want to think back of those concepts or I want to try to interpret or see them from a different point of view, maybe I will. But in introductory stats when they

are just learning I believe that this is very confusing. I don't know if I would give this question in class. Normally when you give a question in class, you should have an answer, a good answer. So I don't believe I will have a good enough answer.

Interviewer: So that your students would have a clear path?

Sandy: Yeah, right. I believe that's what I don't like. Clearly what I don't like is for students to go back with fuzzy concepts, not giving the concepts straight in their heads. I believe that's not good. If it's a literature class where you go and start reading the poetry and you say what you think. Everyone can think different ways or how you feel about it. But it's not. It's a math class basically. Statistics is math, and I expect to give straighter answers.

In this excerpt Sandy argued that this task is not appropriate as a learning tool because it is not a straightforward question with a straightforward answer. She indicated that she prefers learning in a 'straightforward' manner. Although Sandy used the noun "concept" (in the first highlighted excerpt) to describe what was being learned, her discussion is more consistent with learning a "procedure" because her conversation around the term suggested learning a systematic approach rather than an abstract idea. The second time that Sandy used the noun "concept", it is more consistent with the actual definition of the word, where she indicated she might view the topic from a different vantage point, but she seemed to suggest this only after the procedure is taught.

It is also worth noting that a teacher-centered view of teaching and learning radiates through Sandy's response – specifically in the second highlighted utterance, where she discusses the importance of the teacher having a good answer. The metaphor that emerges in this utterance is teacher as authority. That is, the students need to look to the teacher to have an exact answer so if Sandy does not have an exact

answer, then it is not a good problem for her class. Sandy's response also makes me wonder if she conceives of the act of statistical inquiry as black and white, in that every problem has an optimal process that leads to the "correct" solution. Certainly her comment about having a good answer coupled with other comments she made throughout the interviews, like her comments to the Mystery Mixture Task in the following paragraph, for instance, suggest a black and white view of statistical inquiry.

During the Mystery Mixture Task, Sandy again indicated that she is confident in applying formulas and getting correct answers, but she does not like questions where there is no obvious path to follow for a solution. In the following exchange, I asked Sandy to provide an informal confidence interval for the proportion of red candies in the jar after she refused to provide a point estimate.

Interview 1: Mystery Mixture Task:

> **Interviewer:** What if instead of trying to predict the exact amount you were just trying to come up with a confidence interval of like the number of red is between such and such that you would feel pretty confident you would capture it?
> **Sandy:** I cannot do that.
> **Interviewer:** No?
> **Sandy:** No and I don't like to do that.
> **Interviewer:** Because it's too iffy?
> **Sandy:** Exactly. I don't like. I, so how I approach problems is give me the information I will apply the formulas. I will give you the answer. And I'm confident of doing that.

The highlighted excerpt reveals Sandy's personal preference for procedural knowledge and her own confidence in working out procedural problems. Indeed, during the interview when I asked Sandy more open-ended statistical questions or estimation

questions she did not enjoy solving them, and struggled with applying her knowledge of statistics to non-routine problems.

On the one hand, Amanda and Sandy espoused good teaching practices as imparting both procedural and conceptual knowledge. Yet, both Amanda and Sandy struggled more with the conceptual tasks (Real/Fake, Mystery Mixture, Unusual Sample Tasks). They appeared to have stronger procedural knowledge than conceptual knowledge (although this was not the case for all tasks) and their discussions of their teaching practices suggested more of an emphasis on procedural knowledge. This finding is consistent with Eisenhart et al. (1993). On the other hand, when the TAs had an opportunity to explore statistical ideas from a more conceptual point of view, they did so. For example, when I asked Sandy to describe the term standard deviation she provided both a procedural and conceptual explanation.

Interview 3: Terminology:

> **Sandy:** Okay. So I have an example for that, so standard deviation. Let's suppose that you did, ahh, let's suppose you did regression. $y=a+bx$ and this is weight [***Sandy points to the y-variable***] based on what? Something, an x-value. And then you compute the mean squared error and you take the square root of that, which is the standard deviation. So how can you interpret that? You say on average your prediction, let's suppose that you get 4lbs, you say on average your prediction for weight is 4 lbs off. That's how you interpret.…So basically the standard deviation is telling you how far away from the mean value, the mean weight, you are in your calculation.

In the first part of the highlighted utterance, Sandy is focused on a value for standard deviation, which is computed in a particular manner, but as she continued, she provided a strong interpretation of the underlying image of the term standard

deviation. I asked Sandy if that is how she explained standard deviation to her students, but as the following excerpt reveals, this image of standard deviation only recently became part of her repertoire.

Interview 3: Terminology:

> **Interviewer:** Is that how you've explained it to your students before? **Sandy:** No ahh, it just came to my mind right now because I'm taking this class with 'Dr. J' and I had the exam yesterday. And yesterday I was looking through about what I did. You know I didn't take before 243, 244 [*undergraduate introductory statistics*]. And I believe I told you before that not always I realize exactly how to explain some concepts. And it was good for me to go to see how he explains some concepts, and standard deviation. Somehow it's a little bit hard for students to grasp because when you introduce standard deviation you introduce variance. And in fact you show it like that [*Sandy writes the formula for variance on her sheet*]. You know, you show this formula and then they are, they get somehow stuck, you know with the fact that variance is that formula, that complicated formula you know. And they don't really, so somehow they focus on the formula and they don't get the right interpretation of standard deviation which is square root of that [*points to the variance formula*], which is telling you how far away your observations are from the mean.

The graduate statistics course for non-statistics majors that Sandy was enrolled in at the time of our interviews appeared to be a strong influence on her statistical knowledge for teaching. Prior to her new idea for discussing standard deviation, Sandy indicated that she introduced standard deviation strictly procedurally, by showing the formula for variance and then taking the square root of that formula. Once Sandy broadened her own conceptual knowledge of standard deviation, and was exposed to a new model for introducing the topic by an expert professor, Sandy revised her teaching repertoire. Her modes of instruction for developing procedural and conceptual knowledge in students were enriched. Sandy's discussion about the image

of repeating the sampling process in relation to confidence intervals, mentioned in Section 6.2, is another example of how this class influenced Sandy's statistical knowledge for teaching. Sandy also suggested that this course helped her view other introductory statistics ideas from a new perspective as well.

6.4 Conclusions

In this chapter I presented a model of the necessary statistical knowledge for teaching – *statistical literacy*, *statistical thinking*, and *knowledge of content and students*. My primary contention is that TAs are in a better position to support student learning in statistics if they have strong subject matter knowledge, including procedural and conceptual knowledge, as well as the "big ideas" related to different statistical concepts and the connections among concepts. That is, TAs need well-developed statistical literacy and thinking skills and should be familiar with common conceptual hurdles. This conjecture is not novel, and has been made by other mathematics education researchers in other areas of the mathematics education curriculum – Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Ball & McDiarmid, 1990; Hill, Rowan & Ball, 2005), and Eisenhart et al. (1993), for example.

Although the TAs in this study showed evidence of having strong statistical content knowledge, each of the TAs had limitations and/or gaps in their knowledge. Key limitations in TAs' statistical reasoning include:

- Ability to reason with experimental data sets
  - Thinking about variability in experimental data

- o Coordinating multiple aspects of the distribution
  - Maintaining an image of a distribution of sample statistics in relation to sampling and statistical inference problems
  - Recognizing the relationship between sample size and sample variability

TAs' content knowledge limitations and/or gaps are likely to translate into content knowledge limitations in students because TAs are not prepared to address these key concepts in instruction. It is likely that TAs can avoid addressing certain difficulties they might have with the statistics content by not using open-ended statistical tasks or non-routine problems in the curriculum. TAs can teach the statistical procedures for routine problems without much difficulty, but does that count as quality teaching?

Given that these TAs experienced some of the same key difficulties as middle, secondary, and tertiary students, I argue that these concepts are not trivial and need to be more explicit in instruction. That is, statistics courses should emphasize reasoning with experimental data, a consideration of variation and sampling distributions throughout the curriculum. In addition, statistics teaching assistants and other statistics graduate students may benefit from an introductory statistics course prior to taking, or concurrent with taking, their first graduate mathematical statistics course. For instance, the graduate statistics course for non-statistics majors, in which Sandy was enrolled, seemed to positively impact her statistical thinking and subsequently her teaching. Perhaps taking or assisting an expert professor with a condensed calculus-based

introductory probability and statistics course would provide TAs an opportunity to improve their statistical literacy and thinking skills. In fact, the next logical step for research in this area could be an investigation into whether taking or assisting with such a course would improve TAs' statistical knowledge for teaching and positively impact undergraduate student achievement.

The second piece of this chapter consisted of a discussion of TAs' beliefs about teaching and learning. It was not my intention to study TAs' beliefs, but due to some of the methodology of certain open-ended interview tasks, TAs discussed their beliefs about how students learn. It is interesting that these discussions tended to also reveal how these TAs preferred their own learning to be structured and how they found they learned best from their own professors. For example, Amanda tended to learn a particular concept as a result of repeating a procedure over and over again; as a result, she tended to equate the long struggle over learning the procedures and the reasons for each step as what led to her eventual conceptual understanding of a topic. It seemed that this led Amanda to believe that students need to learn the procedures first, which is mostly all we can expect from them at an introductory level.

My ability to draw conclusions about TAs' beliefs and the impact of those beliefs on their teaching is certainly limited in the sense that my research tasks were not aimed at identifying beliefs and I did not observe TAs teaching in order to confirm or disconfirm the information on beliefs gathered from the interviews. The information on TA beliefs was an added bonus in this study. However, since it was not the primary research focus and sporadically surfaced during the interviews, there is not substantial

evidence on TAs' beliefs about teaching and learning from which to support an interpretation of how those beliefs specifically influenced their thinking about teaching and learning. The information TAs did provide is useful for thinking about how this information could be used to frame future studies on TAs' beliefs and knowledge for teaching statistics. Certainly an entire study could be devoted to TAs' beliefs about how students learn statistics concepts and procedures. I turn now to the final conclusions of my study – the implications this study has for undergraduate and graduate statistics education, directions for future research, and the limitations of my study.

CHAPTER 7

CONCLUSIONS

This concluding chapter highlights the central findings on TAs' statistical knowledge for teaching sampling concepts. This chapter is presented in four sections. In Section 7.1, I discuss the central findings of this research study. In Section 7.2, I discuss the study's contributions and implications. In Section 7.3, I discuss the study's limitations. In Section 7.4, I conclude the chapter with a discussion that points to potential relevant future research.

7.1 Central Findings

In Chapters 1 and 2, I discussed the importance of sampling in the development of statistical literacy and statistical thinking skills. In addition, I argued that in order to improve undergraduate statistics education, statistics educators needed to begin an investigation into TAs' statistical knowledge for teaching, because TAs play an integral role in undergraduate education. In order to contribute to the research base on the teaching of undergraduate statistics, the goal of this dissertation study, as described in Chapter 1, was to investigate and characterize TAs' statistical knowledge for teaching sampling concepts. Specifically, my aim was to better understand: (1) how TAs reason about sampling processes; (2) TAs' understandings of the relationships between sampling and probability, and sampling and statistical inference; and, (3) how TAs think about teaching and student learning in the context of sampling.

In order to achieve my research goals, I conducted a thorough review of the existing research literature on statistical literacy, thinking and reasoning, as well as teacher knowledge. This background literature provided the following structure for my study: (a) models of students' and teachers' reasoning about sampling processes and how those processes relate to statistical inference; (b) models for the types of reasoning necessary for statistical literacy and statistical thinking skills – both of which are necessary for statisticians and teachers of statistics; and, (c) methods for constructing models of reasoning and assessing teacher knowledge of content and students. The details of this structure were meticulously discussed in Chapters 2 and 3.

The results of my research were parsed out into three chapters – one for each of the salient findings from this study. I review those findings here, beginning with Chapter 4.

7.1.1 Chapter 4: Tensions TAs experienced between theoretical models and experimental data

In Chapter 4, I provided a conceptual analysis and framework for reasoning about empirical sampling distributions from a series of sampling tasks – the Prediction, Real/Fake, and Mystery Mixture Tasks. The conceptual framework of Shaughnessy et al. (2004a&b, 2005) provided the initial foundation for my examination; however, I refined and modified this framework in order to make it applicable and useful as a means for understanding TAs' reasoning about sampling processes. The end product enabled me to discuss what constitutes a coherent and robust understanding of

sampling processes in the context of reasoning about empirical sampling distributions produced from a sampling experiment, and the ways in which this reasoning might develop from less sophisticated reasoning.

In prior research using the same tasks, Shaughnessy et al. (2004a&b, 2005) observed that middle and secondary school students struggled to coordinate multiple attributes of a distribution. Furthermore, they often focused on individual data points and modal values to reason about, and draw conclusions from, the graphs of the empirical sampling distributions. There were exceptions to Shaughnessy et al.'s findings; specifically, many of the students that participated in the teaching episodes were able to coordinate multiple attributes of the distribution in their arguments. There is compelling evidence from the survey and interview data from this study, presented in detail in Chapter 4, that TAs experienced similar difficulty in coordinating multiple attributes of the experimental sampling distributions as they reasoned about the different tasks. This is a surprising and rather unexpected result.

To be clear, the TAs in this study reasoned about the Prediction Task with a higher level of sophistication than what was observed in the middle and secondary students in the study conducted by Shaughnessy et al. (2004a&b, 2005). The majority of survey participants and all of the interview participants used either a formal or informal probability distribution argument to make their predictions for the Prediction Task. However, when confronted with experimental data from a similar sampling context, as in the Real/Fake and Mystery Mixture Tasks, TAs experienced considerable difficulty applying more sophisticated statistical reasoning. TAs tended to revert back to single

attributes such as the shape of the graph or the modal values to make statistical

decisions. In large part, TAs' difficulties reasoning about the experimental sampling

distributions appeared to be rooted in their tensions over how much variability to

expect in the experiment. TAs had a tendency to focus on the variability in the

frequencies of the bars in the histograms of experimental sampling distributions, rather

than statistical variation. TAs also experienced difficulty deciding how to use all four

sampling distributions to better determine the population parameter during the

Mystery Mixture Task. Specifically, Amanda, Sandy, Sam, and Joe experienced

difficulty estimating the means/medians of graphical distributions of data and tended

to rely on the modal values of one or more graphs in making their predictions.

In Chapter 4, I suggested that TAs did not have a strong sense for how much

variability to expect from sample to sample or from experimental sampling

distribution to experimental sampling distribution. In fact, it appeared that several TAs

expected much more variability than is likely to occur. That is, TAs did not think

about placing bounds on variability in the experiment. Saldanha and Thompson (2003)

alluded to this idea of bounded variability in their characterizations of K-12 students'

reasoning about sampling problems. Rubin et al. (1991) also alluded to the idea of

bounded variability in their discussion of a spectrum of student reasoning ranging

from overly focused on sample representativeness to overly focused on sample

variability. In this study, bounded variability served as a useful explanatory construct

for discussing TAs' reasoning in experimental sampling situations. Bounded

variability as an explanatory construct could serve as a useful tool in future studies or

in teaching experiments as a means for discussing TAs' reasoning about variability in different sampling situations.

Maintaining a sense of bounded variability is fundamentally important in making decisions about experimental data. Without an understanding of expected variability within and between samples and experimental sampling distributions TAs will likely experience difficulty applying their knowledge of statistical variation to experimental data sets, and distinguishing between more likely and less likely outcomes. This suggests that TAs may need more experiences working with experimental data sets to gain some intuition for the expected variability from sample to sample or from sampling distribution to sampling distribution. In addition, working with experimental data may provide TAs an opportunity to apply the statistical knowledge they are learning in their graduate course work to experimental situations.

7.1.2 Chapter 5: TAs' knowledge of sampling and statistical inference

In Chapter 5, I provided a conceptual analysis and framework for reasoning about sampling and statistical inference in the context of two specific tasks – the Unusual Sample Task, and the Gallup Poll Task, respectively. The prior research of Kahneman and Tversky (1972), Konold (1989), and Watson and Moritz (2000a) provided a basis for investigating TAs' knowledge of the role between sampling variability and sample size. The conceptual framework of Liu and Thompson (2005) provided the initial foundation for my examination of TAs' understanding of statistical inference. However, I refined and modified these frameworks in order to make them applicable

and useful as a means for understanding TAs' reasoning about sampling and the role that sampling processes play in statistical inference. The end product enabled me to discuss what constitutes a coherent and robust understanding of sampling processes and statistical inference, and the ways in which TAs' interpretations of these tasks influences their solution to the problem.

In tasks comparable to the Unusual Sample Task, Kahneman and Tversky (1972), Konold (1989) and Watson and Moritz (2000a) found that students from elementary school through college struggle to understand the relationship between sample size and sample variability. Often students believe that large samples are likely to have the same amount of variability as small samples. Konold suggested that this is a result of the way in which students interpret the question being asked. Konold found evidence to support the conjecture that students often believe they are being asked to determine whether the large or small sample *will* be more unusual. From this perspective, students often reason that it is not possible to determine whether a large or small sample is more likely to be unusual, since a possibility exists for either sample to produce unusual outcomes.

Although the survey results indicated that most TAs reasoned about the Unusual Sample Task from a perspective of likelihood and used the Law of Large Numbers as support for their reasoning, it is difficult to determine whether or not these TAs in fact had a long-term relative frequency perspective of probability, because of the brevity of their responses. However, a small percentage of survey participants did indicate that both samples were equally likely to be unusual, and provided responses consistent

with Konold's (1989) outcome approach. The interviews provided stronger evidence that at least two of the TAs in this study, Amanda and Sandy, reasoned by the outcome approach on the Unusual Sample Task.

In addition, three TAs, Andy, Joe, and Sam, also appeared to reason using the outcome approach on the Gallup Poll Task. That is, these TAs suggested that the confidence level was in reference to the individual samples, rather than the distribution of sample statistics. Furthermore, these TAs reasoned differently in different contexts. The interview data strongly suggests that these TAs did not have a robust understanding of the hypothetical repetition of the sampling process, that this process builds a distribution of sample statistics, and/or how this concept underscores the foundation of statistical inference.
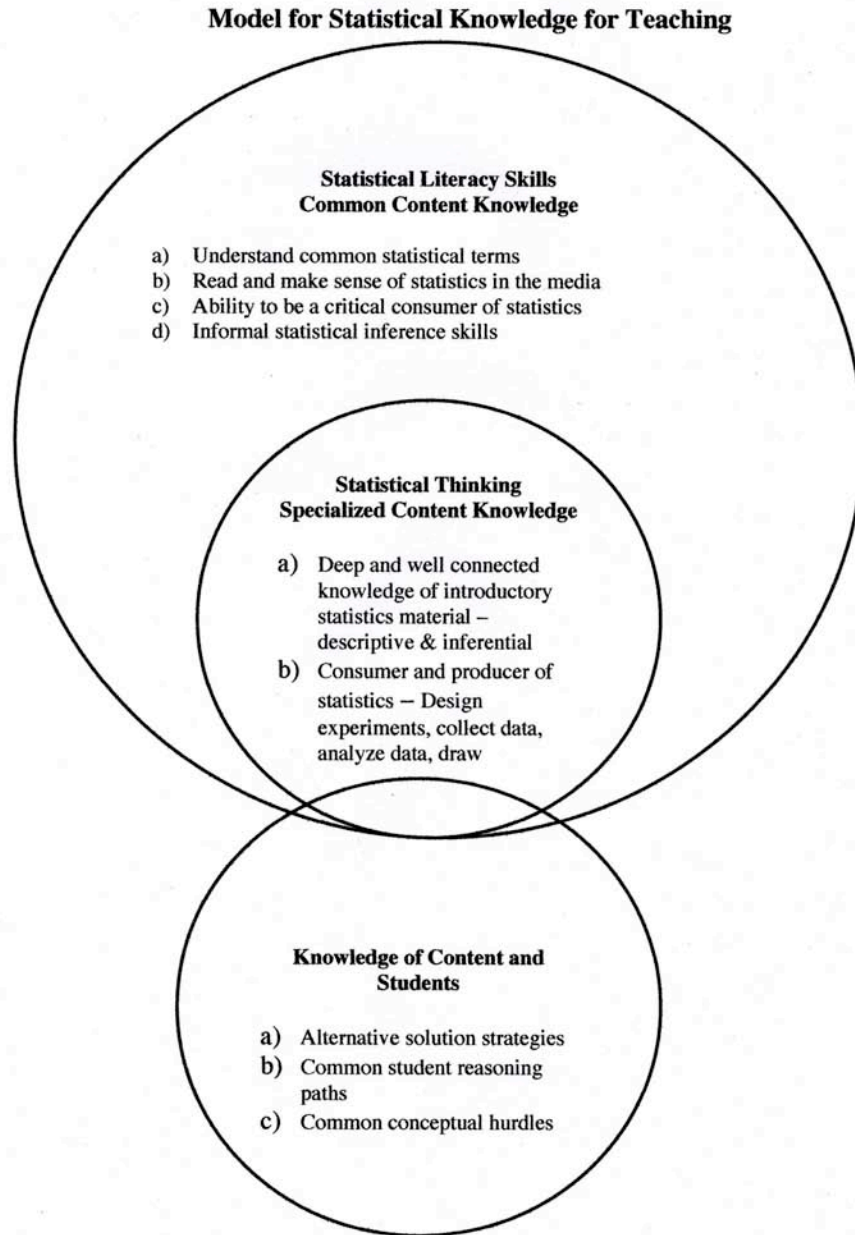
I suggest, in accordance with Saldanha and Thompson (2003, in press), that the concept of repeated sampling is an elemental image necessary for coherent understandings of statistical inference. The results of Chapter 5 suggest that the concept of repeated sampling is foundational and must be developed in statistics instruction in the middle, secondary and tertiary grades in order to facilitate students' learning of statistical inference in a coherent manner. That repeated sampling is a foundational aspect of statistical inference is akin to the foundational concept of partitioning for developing a coherent understanding of fractions, or the foundational concept of rates of change for developing a coherent understanding of differential calculus.

7.1.3 Chapter 6: TAs' Statistical Knowledge for Teaching Sampling

In Chapter 6, I proposed a model for describing the types and quality of necessary statistical knowledge for teaching sampling. This model was built by fusing the constructs of statistical literacy, thinking and reasoning from the stochastics education research with current research on teacher knowledge, most notably with addition of the construct of mathematical knowledge for teaching from the work of Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Ball & McDiarmid, 1990; Hill, Rowan & Ball, 2005), and the work of Eisenhart et al. (1993) on teachers' procedural and conceptual knowledge. The visual representation from this model is shown again in Figure 47. This model was then applied toward my examination of the interview participants' statistical knowledge for teaching.

Figure 47: Model of Statistical Knowledge for Teaching

Model for Statistical Knowledge for Teaching

Statistical Literacy Skills
Common Content Knowledge

a) Understand common statistical terms
b) Read and make sense of statistics in the media
c) Ability to be a critical consumer of statistics
d) Informal statistical inference skills

Statistical Thinking
Specialized Content Knowledge

a) Deep and well connected knowledge of introductory statistics material – descriptive & inferential
b) Consumer and producer of statistics – Design experiments, collect data, analyze data, draw

Knowledge of Content and Students

a) Alternative solution strategies
b) Common student reasoning paths
c) Common conceptual hurdles

The findings related in Chapters 4 and 5 reveal that TAs may not have as robust a

content knowledge as necessary for teaching introductory statistics. Concepts may not

be well connected and/or TAs may have non-normative views about expected variability within experimental data and/or how to interpret sampling and statistical inference problems. TAs' difficulties are likely to transfer to student difficulties and may limit TAs' ability to develop students' statistical literacy and thinking skills. In addition, TAs did not have strong knowledge of content and students, including the types of difficulties students are likely to experience with certain topics and the developmental stages they are likely to pass through. That TAs lack knowledge of content and students is not surprising in that TAs are not necessarily studying issues in statistics education and have not taught the introductory statistics enough times to be familiar with particular student difficulties. Yet, it may be appropriate for TAs to have some exposure to research on students' statistical reasoning in their orientation courses in order for them to develop strong teaching skills.

7.2 Contributions and Implications

There are two key contributions that this study offers to the statistics education community. First, the conceptual analysis and frameworks that emerged for how TAs reasoned about sampling processes constitutes a significant contribution to the statistics education community for the following reasons: (1) this research extends aspects of the research literature on students' and teachers' conceptions of sampling; (2) this is the first study on TAs' statistical knowledge for teaching and fills a void in the research literature; and, (3) this study has the potential to improve undergraduate statistics education by improving TAs' statistical knowledge for teaching. In Chapter 1, I identified the importance of making improvements in undergraduate statistics

education and the importance of considering the role of TAs in undergraduate statistics education. This study has made the first contribution in considering the role of TAs in undergraduate statistics education.

The frameworks, presented in Chapters 4, 5 and 6, advance our understanding of reasoning about sampling processes and what constitutes strength and depth of understanding. For example, I demonstrated that the elemental images of bounded variability and repeated sampling appear foundational to developing a coherent understanding of statistical inference. Thus, these concepts provide explanatory constructs that need to be explicitly targeted in instruction. The frameworks also indicate alternative, non-normative ways of reasoning that TAs use for understanding sampling and the relationship between sampling and statistical inference. Taken together, the frameworks constitute a tool for thinking about how TAs reason and how to support the development of strong conceptions of sampling and statistical inference. In addition, that TAs experienced certain difficulties in reasoning about sampling and statistical inferences is surprising and suggests that sampling concepts are not trivial. Reasoning about experimental data and, in particular, resolving the tensions between representativeness and variability, and understanding how a distribution of sample statistics can be applied to numerous concepts in statistics, all prove to be rather challenging endeavors, even for people with considerable statistical experiences.

A second significant contribution this study offers is the framework that emerged for thinking about TAs' statistical knowledge for teaching. Statistics educators have called for reform and improvement in K-12 as well as undergraduate statistics

education (Shaughnessy, 2007) in order to promote statistical literacy and thinking skills among all students. Achieving this goal requires research efforts aimed at better understanding student thinking, and research efforts aimed at better understanding the statistical knowledge necessary for effective teaching. My framework points to characteristics such as content knowledge and knowledge of students' statistical development that would enable the emergence of profound statistical knowledge for teaching skills. Specifically, I characterized: (a) normative ways of thinking about sampling that are necessary for teaching sampling topics; (b) non-normative ways in which TAs may conceive of sampling processes or their relation to statistical inference; (c) ways in which TAs thought about student learning; and, (d) whether or not TAs' knowledge of content and students mapped to the statistics education community's knowledge of content and students.

In light of the findings from Chapters 4 and 5, compelling evidence exists that suggests particular concepts which need to be addressed in TAs' graduate school experience in order to enable the maturation of their statistical thinking skills. Most of the TAs in this study had taken multiple graduate statistics courses, and yet they appeared to experience difficulty applying more sophisticated forms of statistical reasoning in novel settings. TAs reverted back to statistically naïve forms of reasoning, such as shape and modal values, in attempting to make decisions based on experimental information. In addition, TAs did not appear to have a robust knowledge of the importance of a distribution of sample statistics in sampling and statistical inference topics. Thus, perhaps more abstract, graduate level work is not sufficient for

349

enabling the evolution of TAs' understandings of sampling topics. Specifically, I argue that TAs could benefit from the opportunity to make conjectures about sampling tasks, perform sampling experiments to test those conjectures, and then reflect upon the results. These hands-on experiments may enable a deeper, richer development of TAs' reasoning about variability of experimental data. These experiences may lead to a deeper understanding of the effects of sample size on sample variability, the applicability of a distribution of sample statistics to the foundations of statistical inference, and the application of their formal statistical training to novel experimental situations.

The frameworks on TAs' reasoning about sampling and on the necessary knowledge for teaching sampling concepts may also provide insights for the construction of professional development or mentoring opportunities for TAs teaching introductory statistics courses. One strategy for professional development gleaned from this dissertation study would include engaging TAs in sampling tasks like the ones used in this research in order to:

- Strengthen TAs' ability to apply their knowledge of distribution to experimental situations.

- Provide TAs the opportunity to grapple with and strengthen their understanding of variability within experimental situations.

- Provide TAs the opportunity to consider on a deeper level their interpretations of probability in different contexts and to make connections

between probability and sampling concepts, and probability and statistical inference.

- Provide TAs the opportunity to explore models of student reasoning and common conceptual hurdles to understanding sampling topics as a basis for effecting change in TAs' teaching practices.

7.3 Limitations

Of course, all empirical studies have limitations. In hindsight there are several limitations to this study. First, the ability to generalize my results to a larger population of TAs is limited. My participants could not be randomly selected and comprised a convenience sample. In addition, I was unable to collect information about the larger population of TAs from each of the universities that comprised my survey pool. For example, I could not obtain information on the age, gender, English as a second language, or statistics backgrounds of all the mathematics and statistics TAs from the 18 universities that participated in my study. This information would have been useful in order to gauge whether or not the subset of TAs that participated in my study from each university were representative of the larger group of TAs at each university. Also, in some cases I had only 2 participants from one university and 12 participants from another university, making it difficult to compare across universities. However, given that the limitations I observed in TAs' reasoning have been observed in K-12 students and K-12 teachers, I highly suspect that the difficulties experienced by these TAs are not unusual.

Second, I was unable to test my conjectures about using sampling tasks like the ones in the present study as a means for improving TAs' statistical content knowledge. Thus, there is no information on whether engaging TAs in these types of hands-on tasks would improve their statistical knowledge for teaching. Third, I was unable to test my conjectures about the necessary statistical knowledge for effective teaching on student achievement. The research of Ball and her colleagues (Ball, Hill & Bass, 2005; Ball & Bass, 2003; Ball, Lubienski & Mewborn, 2001; Ball & McDiarmid, 1990; Hill, Rowan & Ball, 2005) points to higher student achievement among students whose teachers have higher scores on the mathematical knowledge for teaching assessment created by Ball and her colleagues. Unfortunately, there is no data available on whether my suggested improvements for TAs' statistical knowledge for teaching would translate into gains in student achievement.

Finally, the methodology of providing hypothetical student responses proved successful in eliciting TAs' content knowledge, and their knowledge of content and students. Unfortunately, in certain tasks I asked rather open-ended questions about the types of difficulties TAs expected students to have. These open-ended questions provided interesting information on TAs' beliefs about teaching and how students learn, but it did not access TAs' knowledge of student development in the same way that the hypothetical student responses did. In retrospect I would have developed more hypothetical student responses for each task. I believe this would have provided a clearer picture of TAs' knowledge of content and students.

7.4 Suggestions for Future Research

There are many directions for future research that could enable a greater understanding of TAs' statistical knowledge for teaching, and what features of their statistical knowledge for teaching could result in gains in student achievement. An obvious next step for this research would be to conduct a seminar or teaching experiment using the same research tasks to study the effects of such an experience on TAs' statistical reasoning. There is some evidence from the interview data that suggests this could result in positive outcomes in TA reasoning. For example, there were times during the interviews when TAs mentioned that they had never " had to pick apart " how they thought about certain ideas which were raised during the interviews. The interview participants often said things like "wait, I'm putting all the pieces together", or "I've never thought about that before". These comments suggest that the tasks challenged TAs to make sense of subtle concepts that they had perhaps taken for granted in the past. It also suggests that TAs were engaged in a learning experience during the interview conversations.

A natural next step in this research would be to develop a seminar for TAs that would engage them in sampling tasks, activities and conversations around student thinking and learning in this area. In addition, assessments could be designed that would measure TAs' statistical knowledge for teaching. A large comparison study between seminar participants and non-participants and their students' achievement could be conducted in order to see if the seminar was effective in improving TAs' statistical knowledge for teaching, and subsequently students' statistical knowledge.

REFERENCES

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147-168). Dordrecht, The Netherlands: Kluwer.

Ball, D. L. (2005). Mathematics teaching and learning to teach project. *American Educational Research Association: Annual Meeting Montreal*. Retrieved June 2005 from http://www-personal.umich.edu/~dball/

Ball, D. L., & Bass, H. (2003). Toward a practice based theory of mathematical knowledge for teaching. In B. Davis & E. Simmt (Eds.), *Proceeding of the 2002 Annual Meeting of the Canadian Mathematics Education Study Group,* (pp. 3-14). Edmonton, AB: CMESG/GCEDM.

Ball, D. L., Hill, H. C., Bass, H. (2005). Knowing mathematics for teaching: Who knows math well enough to teach third grade and how can we decide? *American Educator*, pp. 14-46.

Ball, D. L., Lubienski, S.T., & Mewborne, D.S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp.433-456). New York: Macmillan.

Ball, D. L., & McDiarmid, G. W. (1990). The subject-matter preparation of teachers. In R. W. Houston, M. Haberman & J. Sikula (Eds.), *The handbook of research on teacher education: A project of the association of teacher educators* (pp. 437-449). New York: Macmillan.

Batanero, C., Tauber, L.M., & Sanchez, V. (2004). Students' reasoning about the normal distribution. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 257-276). Dordrecht, The Netherlands: Kluwer.

Begle, E. G. (1979). *Critical variables in mathematics education: Findings from a survey of the empirical literature*. Washington, DC: Mathematical Association of America and National Council of Teachers of Mathematics.

Belnap, J.K. (2005). *Putting TAs into context: Understanding the graduate mathematics teaching assistant*. Doctoral dissertation, University of Arizona.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer.

Callingham, R.A., Watson, J.M., Collis, K.F., & Moritz, J.B. (1995). Teacher attitudes towards chance and data. In B. Atweh & S. Flavel (Eds.), *Proceedings of the Eighteenth Annual Conference of the Mathematics Education Research Group of Australasia* (pp.143-150). Darwin, NT: Mathematics Education Research Group of Australasia.

Canada, D. (2004). *Pre-service elementary teachers conceptions of variability.* Unpublished doctoral dissertation, Portland State University, Portland, OR.

Carpenter, T.P., Fennema, E., Peterson, P.L., Chiang, C., Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26*, 499-531.

Carpenter, T.P., Fennema, E., Peterson, P.L., & Carey, D.A. (1988). Teachers' pedagogical content knowledge of students' problem solving in elementary arithmetic. *Journal for Research in Mathematics Education, 19(5),* 385-401.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-324). Dordrecht, The Netherlands: Kluwer Academic.

Clark, C. & Peterson, P. (1986). Teachers' though processes. In Wittrock, M.C. (Ed.) *Handbook of Research on Teaching, 3rd Edition.* New York: Macmillan.

Cobb, G. W. (1999). Individual and collective mathematics development: The case of statistical analysis. *Mathematical Thinking and Learning, 1*, 5-44.

Cobb, G.W., & Moore, D.S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104(9)*, 801-823.

Cobb, G.W. (1998). The objective-format question in statistics: Dead horse, old bath water, or overlooked baby? *Paper presented at the annual meeting of the American Educational Research Association*, San Diego, CA.

Cobb, G.W. (1993). Reconsidering statistics education: A national science foundation conference. *Journal of Statistics Education, 1(1)*. Retrieved May 25, 2006, from http://www.amstat.org/publications/jse/v1n1/cobb.html.

Davis, R.B., Maher, C.A. & Noddings, N. (1990) Introduction: Constructivist views on the teaching and learning of mathematics [*Monograph*]. *Journal for Research in Mathematics Education*, *4*, pp. 1-3.

DeFranco, T.C., & McGivney-Burelle, J. (2001). *The beliefs and instructional practices of mathematics teaching assistants participating in a mathematics pedagogy course*. Paper Presented at the 23rd Annual Conference of Psychology of Mathematics Education – North America, Snowbird, Utah.

Devore, J. L. (2004). *Probability and statistics for engineering and the sciences* (6th editions). California: Brooks/Cole – Thompson Learning.

Eisenhart, M., Borko, H., Underhill, R., Brown, C., Jones, D., & Agard, P. (1993). Conceptual knowledge falls through the cracks: Complexities of learning to teach mathematics for understanding. *Journal for Research in Mathematics Education, 24*(1), pp. 8-40.

Even, R. (1993). Subject-matter knowledge and pedagogical content knowledge: Prospective secondary teachers and the function concept. *Journal for Research in Mathematics Education, 24, 2*, pp.94-116.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education, 24*, 4, pp. 403-434.

Fennema, E., & Franke, M. L. (1992). Teachers' knowledge and its impact. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp.147-164). New York: Macmillan.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*, 1-25.

Gal, I. (2003). Expanding conceptions of statistical literacy: An analysis of products from statistics agencies. *Statistics Education Research Journal, 2*(1), 3-21.

Gal, I. (2004). Statistical literacy: Meanings, components, responsibilities. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 47-78). Dordrecht, The Netherlands: Kluwer Academic.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology Volume 2* (pp. 82-115). John Wiley & Sons Ltd.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103*, 3, pp. 592-596.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine.

Heid, M. K., Perkinson, D., Peters, S. A., & Fratto, C.L. (2005). Making and managing distinctions – the case of sampling distributions. In Lloyd, G. M., Wilson, M., Wilkins, J. L. M., & Behm, S. L. (Eds.). *Proceedings of the 27th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education.*

Hiebert, J. (1986). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hill, H. C., Schilling, S. G., Ball, D. L. (2004). Developing measures of teachers' mathematical knowledge for teaching. *The Elementary School Journal, 105*, 1, pp. 11-29.

Hill, H.C., Rowan, B., & Ball, D.L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42(2),* 371-406.

Hogg, R. V., & Craig, A. T. (1995). *Introduction to mathematical statistics* (5th ed). New Jersey: Prentice Hall.

Jacobs, V. R. (1997, April). *Children's understanding of sampling in surveys*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Kagan, D.M. (1992). Implications of research on teacher belief. *Educational Psychologist, 27(1),* 65-90.

Kahneman, D. & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology, 3,* 430-454.

Koehler, M.S., & Grouws, D.A. (1992). Mathematics teaching practices and their effects. In D.A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 115-126). New York: Macmillan.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6,* 59-98.

Konold, C. (1991). Understanding students' beliefs about probability. In E. von Glasersfeld (Ed.), *Radical Constructivism in Mathematics Education* (pp. 139-156). The Netherlands: Kluwer Academic.

Konold, C., & Pollatsek, A. (2002). Data analysis as a search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*, 259-289.

Konold, C., Pollatsek, A., Well, A., Lohmeier, J. & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal of Research in Mathematics Education, 24(5)*, 392-414.

Liu, Y. (2004). *Teachers' understandings of probability and statistical inference and their implications for professional development*. PhD Dissertation, Vanderbilt University.

Liu, Y. & Thompson, P. (2005). Understandings of margin of error. In Lloyd, G. M., Wilson, M., Wilkins, J. L. M., & Behm, S. L. (Eds.). *Proceedings of the 27ᵗʰ annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*.

Lortie, D. (1975). *Schoolteacher: A sociological study*. Chicago: University of Chicago Press.

Luzter, D.J., Maxwell, J.W., Rodi, S.B. (2000). *Statistical abstract of undergraduate programs in the mathematical sciences in the United States: Fall 2000 CBMS Survey*. Retrieved on 5/17/05 at http://www.ams.org/cbms/cbms2000.html

Maher, C. A. & Alston, A. (1990) Teacher development in mathematics in a constructivist framework [*Monograph*]. *Journal for Research in Mathematics Education*, *4*, pp. 1-3.

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning, and thinking* (pp. 353-374). Dordrecht, The Netherlands: Kluwer Academic.

McClave, J.T., & Sincich, T. (2000). *Statistics* (8th ed.). New Jersey: Prentice Hall.

Mickelson, W., & Heaton, R. (2004). Primary teachers' statistical reasoning about data. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 327-352). Dordrecht, The Netherlands: Kluwer Academic.

Moore, D.S. (2005). Preparing graduate students to teach statistics: Introduction. *The American Statistician, 59(1)*, pp. 1-3.

National Council on Education and the Disciplines. (2001). *Mathematics and democracy: The case for quantitative literacy*. The Woodrow Wilson National Fellowship Foundation.

National Council of Teachers of Mathematics. (2000). *Principals and standards for school mathematics*. National Council of Teachers of Mathematics, Inc., Reston, VA.

Nathan, M. J., & Koedinger, K. R. (2000). An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction, 18(2),* pp. 209-237.

Noddings, N. (1990). Constructivism in mathematics education [*Monograph*]. *Journal for Research in Mathematics Education*, *4*, pp. 7-18.

Pajares, M.F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. Review of Educational Research, 62, 307-332.

Peterson, P.L., Fennema, E., Carpenter, T.P., & Loef, M. (1989). Teachers' pedagogical content beliefs in mathematics. *Cognition and Instruction 6(1)*, 1-40.

Piaget, J. (1971). *Genetic Epistemology*. New York, NY: W. W. Norton & Company.

Pfannkuch, M. (2005). Probability and statistical inference: how can teachers enable learners to make the connection? Graham A. Jones (Ed.). In *Exploring probability in school: challenges for teaching and learning*, 267-294. The Netherlands: Kluwer Academic Publishers.

Pfannkuch, M., & Wild, C.J. (2004). Towards an understanding of statistical thinking. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17-46). Dordrecht, The Netherlands: Kluwer Academic.

Reading, C. & Shaughnessy, J.M. (2000). Student perceptions of variation in a sampling situation. IN T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89-96). Hiroshima, Japan: Hiroshima University.

Reading, C. & Shaughnessy, J.M. (2004). Reasoning about variation. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer Academic.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Vol. 1, pp. 314-319). Voorburg, The Netherlands: International Statistical Institute.

Saldanha, L. & Thompson, P. (in press). Exploring connections between sampling distributions and statistical inference: An analysis of students' engagement and thinking in the context of instruction involving repeated sampling. *International Electronic Journal of Mathematics Education.*

Saldanha, L., & Thompson, P. (2003). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257-270.

Shaughnessy, J.M. (1992). Research on probability and statistics: Reflections and directions. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465-494). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 957 – 1009). Charlotte NC: National Council of Teachers of Mathematics.

Shaughnessy, J.M., & Chance, B. (2005). *Statistical questions from the classroom.* Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J.M., Ciancetta, M., & Canada, D. (2004a). Types of student reasoning on sampling tasks. In M. Johnsen Høines & A. Berit Fuglestad (Eds.). *Proceedings of the 28th meeting of the International Group for Psychology and Mathematics Education* (Vol. 4, pp. 177-184). Bergen, Norway: Bergen University College Press.

Shaughnessy, J.M., Ciancetta, M., Best, K., & Canada, D. (2004b, April). *Students' attention to variability when comparing distributions*. Paper presented at the Research Pre-session of the 82nd annual meeting of the National Council of Teachers of Mathematics, Philadelphia, PA.

Shaughnessy, J.M., Ciancetta, M., Best, K., & Noll, J. (2005, April). *Secondary and middle school students' attention to variability when comparing data sets*. Paper presented at the Research Pre-session of the 83rd annual meeting of the National Council of Teachers of Mathematics, Anaheim, CA.

Shaughnessy, J. M., Watson, J. M., Moritz, J. B., & Reading, C. (1999, April). *School mathematics students' acknowledgement of statistical variation: There's more to life than centers*. Paper presented at the Research Pre-session of the 77th annual meeting of the National Council of Teachers of Mathematics, San Francisco, CA.

Shavelson, R.J., Webb, N.M., & Burstein, L. (1986). Measurement of teaching. In Wittrock, M.C. (Ed.). *Handbook of Research on Teaching, 3rd Edition*. New York: Macmillan.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educational Researcher, 15*, 2, pp. 4-14.

Shulman, L.S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*, 1-22.

Speer, N., Gutmann, T., Murphy, T.J. (2005). Mathematics teaching assistants preparation and development. *College Teaching, 53(2)*, 75-80.

Speer, N. M. (2001). *Connecting beliefs and teaching practices: A study of teaching assistants in collegiate reform calculus courses*. PhD dissertation., University of California, Berkeley.

Tall D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics 12*, pp. 151-169.

Tempelaar, D. (2002). Modeling students' learning of introductory statistics. In B. Phillips (Ed.), *Proceedings of the 6th International Conference on Teaching Statistics*. Cape Town, South Africa.

Thompson, A.G. (1992). Teachers' beliefs and conceptions: A synthesis of the research. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127-146). New York: Macmillan.

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal, 12*, 2, pp. 147-169.

Tversky, A., Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

von Glasersfeld, E. (1990). An exposition of constructivism: Why some like it radical [*Monograph*]. *Journal for Research in Mathematics Education*, *4*, pp. 19-30.

von Glasersfeld, E. (1995). Radical constructivism: A way of knowing and learning. The Falmer Press, Washington D.C.

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88(421)*, 1-8.

Watson, J.M. (2001). Profiling teachers' competence and confidence to teach particular mathematics topics: the case of chance and data. *Journal of Mathematics Teacher Education, 4*, 305-337.

Watson, J. M. (2004). Developing reasoning about samples. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277-294). Dordrecht, The Netherlands: Kluwer Academic.

Watson, J. M. (2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics, 51*, 225-256.

Watson, J. M., & Kelly, B. A. (2004). Expectation versus variation: Students' decision making in a chance environment. *Canadian Journal of Science, Mathematics, and Technology Education, 4*, 371-396.

Watson, J. M., & Moritz, J. B. (2000a). Development of understanding of sampling for statistical literacy. *Journal of Mathematical Behavior, 19*, 109-136.

Watson, J. M., & Moritz, J. B. (2000b). Developing concepts of sampling. *Journal for Research in Mathematics Education, 31*, 44-70.

Watson, J.M., & Moritz, J. B. (1997). Teachers' views of sampling. In N. Scott & H. Hollingsworth (Eds.), *Mathematics Creating the Future* (pp.345-353). Adelaide: Australian Association of Mathematics Teachers, Inc.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223-265.

Winkler, R. (1972). *Introduction to Bayesian inference and decision.* New York: Holt, Rinehart & Winston.

APPENDIX

TASKS

Prediction Task

PREDICTION TASK

A jar contains 1000 candies, 750 are red and 250 are yellow. The candies are mixed well. Suppose that you pull a random sample of 10 candies from the jar, record the number of reds, put the candies back in the jar and mix them up. Suppose you do this 50 times. How many times out of 50 do you think you would get a handful of 10 candies with:

Number of Red Candies in Handfuls of 10	Prediction
0 red candies	
1 red candies	
2 red candies	
3 red candies	
4 red candies	
5 red candies	
6 red candies	
7 red candies	
8 red candies	
9 red candies	
10 red candies	
Total	50

Hypothetical Student Responses for the Prediction Task

The table below shows the predictions of two introductory statistics students

Number of Red Candies in Handfuls of 10	Predictions Student 1	Predictions Student 2
0 red candies	0	1
1 red candies	0	1
2 red candies	0	1
3 red candies	0	1
4 red candies	0	2
5 red candies	5	3
6 red candies	9	4
7 red candies	15	17
8 red candies	18	18
9 red candies	3	1
10 red candies	0	1
TOTAL	50	50

a) In your statistical opinion, Student 1 gave a(n):

 ☐ Reasonable prediction   ☐ Unreasonable prediction

  Explain how you came to this decision.

  If you think this student's prediction is unreasonable explain what the student was thinking that could have resulted in an unreasonable prediction:

b) In your statistical opinion, Student 2 gave a(n):

  ☐ Reasonable prediction   ☐ Unreasonable prediction

  Explain how you came to this decision:

  If you think this student's prediction is unreasonable explain what the student was thinking that could have resulted in an unreasonable prediction:

Real/Fake Task

Real/Fake Task:

A class conducted an experiment, pulling 50 samples of 10 candies from a jar with 750 red and 250 yellows, and graphed the number of reds. However, in this class some of the groups 'cheated' and did not really do the experiment, they just made up a graph. Here are some of the students' graphs from that class. Which graphs do you think are real? Which graphs do you think are made-up? Explain the reasons for your choices.
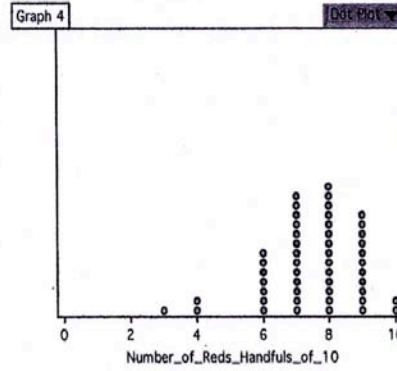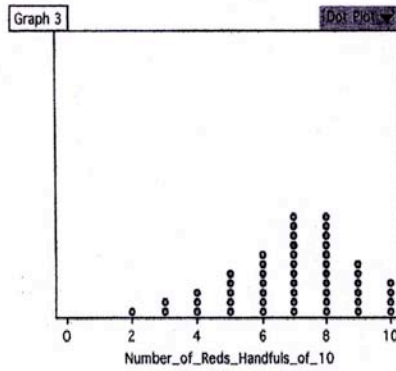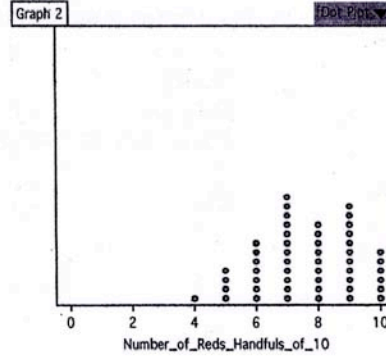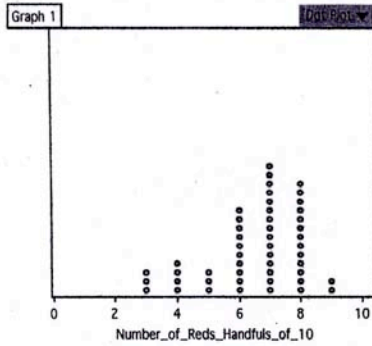
Mystery Mixture Task

Mystery Mixture Task

The graphs below all came from a class that is trying to estimate a mystery mixture of 1000 red and yellow candies in a large jar. They pulled 50 samples of size 10 (recording the number of reds and then replacing and remixing each time). Here are the graphs for the number of reds for four groups from that class.
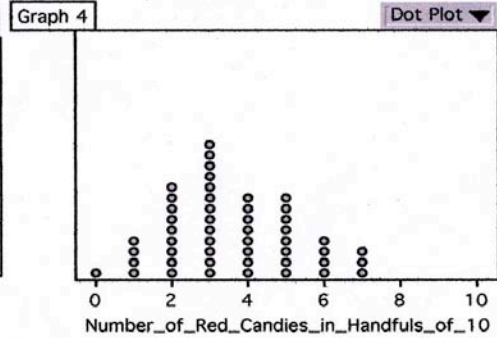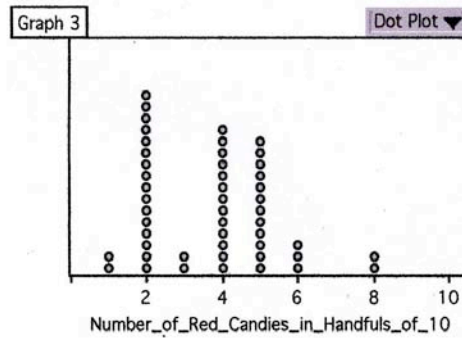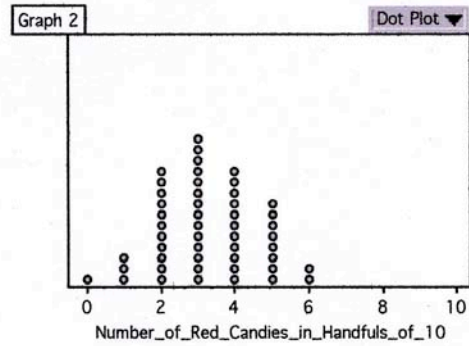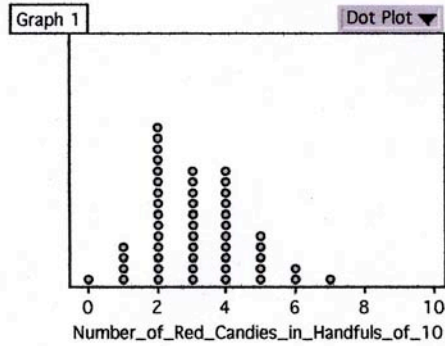
a) What do you think the mixture in the jar might be?
b) Explain why you think this.

The Unusual Sample Task

> ## The Unusual Sample Task
>
> Researchers from Erie County in Buffalo were studying the weight of Grade 5 children. The researchers went to 2 schools: one school was in the center of the city and one school was in the country. Each school had about half girls and half boys. The researchers took a random sample from each school: 50 children from the city school, 20 children from the country school. One of these samples was unusual because it had more than 80% boys. Is the unusual sample more likely to be the large sample of 50 from the city school, the small sample of 20 from the country school, or are both samples equally likely to be the unusual sample?

Gallup Poll Task

> ### Gallup Poll Task
> Your statistics class was discussing a Gallup poll of 500 Oregon voters' opinions regarding the creation of a state sales tax. The poll stated, "…the survey showed that 36% of Oregon voters think a state sales tax is necessary to overcome budget problems". The poll had a margin of error of ± 4%. Discuss the meaning of margin of error in this context.

Gallup Poll Task: Hypothetical Student Interpretations of Margin of Error

Student A says: The margin of error being 4% means that between 32% and 40% of all Oregon voters believe an income tax is necessary.

Student B says: We don't know if the interval 32% to 40% contains the true percentage of voters that believe an income tax is necessary, but if we sample 100 times, about 94 of those times the interval would capture the true percentage of voters.

Student C says: The interval 32% to 40% will be off about 4% of the time, or 4 out of 100 times.

Student D says: If you performed repeated samples of 500 voters, the proportion of voters in favor of sales tax in these samples would fall within the interval 32% to 40%, the majority of the time.

Student E says: I can be 95% sure that all the sample statistics will fall within ± 4% of the unknown population parameter.

Student F says: The interval 36% ± 4% has a high probability (approximately 95%) of being repeated if the sample was repeated.

Gallup Poll Task: Investigating Confidence Level

Hypothetical Student 1: A 95% confidence level means that you can be 95% confident that the particular interval found in the survey captures the population proportion. Do you agree or disagree with this student's interpretation of confidence? Explain.

Hypothetical Student 2: A 95% confidence level means that you are 95% confident in the estimation process. That is, 95% of the time you get good interval estimates that capture the population proportion. Do you agree or disagree with this student's interpretation of confidence? Explain.

What would the confidence level be for this Gallup poll?

How do you interpret confidence level in this context?